БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Н. К. Чертко

МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ЗЕМЛЕУСТРОЙСТВЕ

Рекомендовано Учебно-методическим объединением
по естественнонаучному образованию
в качестве учебно-методического пособия
для студентов учреждения высшего образования,
обучающихся по специальности
1-31 02 01 «География (по направлениям)»,
направление специальности
1-31 02 01-03 «География (геоинформационные системы)»

МИНСК
БГУ
2014

УДК 336

Р е ц е н з е н т ы:
заведующий кафедрой физической географии
УО «Белорусский государственный педагогический университет
имени Максима Танка», кандидат географических наук,
доцент *А. В. Таранчук*;
доктор географических наук *В. С. Хомич*

Чертко, Н. К.

Ч-50 Математические методы в землеустройстве [Электронный ресурс] : учеб.-метод. пособие / Н. К. Чертко. – Минск : БГУ, 2014.
ISBN 978-985-566-045-4.

Рассматриваются основы статистики и математические методы, используемые в землеустроительных исследованиях с целью установления степени сходства или различия и зависимости объектов, их классификации, динамики развития.

УДК 336

ISBN 978-985-566-045-4

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Математические методы – специальная научная и учебная дисциплина, предметом изучения которой являются выборочные совокупности и их оценка. Варианты (результаты наблюдения или эксперимента) в выборочной совокупности не имеют постоянных, одних и тех же исходов. Например, урожай культур, продуктивность растительности в городских ландшафтах меняется ежегодно, а прибыль предприятия – ежемесячно. Однако многие хаотичные явления имеют упорядоченную структуру, поэтому могут иметь конкретные численные оценки. Главное условие для этого – статистическая устойчивость этих явлений, т. е. существование определенных закономерностей, которые можно описать математическими моделями статистически. Ведь большинству природных и экономических явлений свойственна вариабельность (изменение во времени в определенных пределах).

Значительное влияние на развитие математических методов оказали закон больших чисел, открытый Яковом Бернулли (1654–1705), и появление теории вероятности, основы корой разработал французский математик и астроном Пьер Симон Лаплас (1749–1827).

На основе теории вероятности, которая позволяет выявлять определенные тенденции в кажущемся хаосе случайных явлений, появились методы математической статистики. Предметом математических методов стала формально математическая сторона статистического анализа и на ее основе количественная оценка вероятностей различных явлений.

Большинство математических методов универсальны и применяются в различных отраслях деятельности человека. Поэтому многие компьютерные статистические программы не выступают чисто землеустроительными прикладными программами. Иногда выполнение отдельных функций в таких программах не является по сути статистическим.

Все программные средства, которые можно использовать для статистической обработки на персональных компьютерах, можно разделить на:

● Специализированные пакеты, например «Мезозавр» – программа анализа временных рядов. Используется для решения узкого круга задач и специальных методов статистического анализа.

● Статистические методы общего назначения: «Диастат», STADIA, STATGRAPHICS. Они предназначены для широкого круга специалистов различного профиля.

● Табличные процессоры и электронные таблицы QUATROPRO, Excel.

Особенностью любого пакета статистических программ является выдача большого количества информации, которая описывает результат статистического анализа. При недостаточной подготовке пользователь не способен правильно все воспринять и осмыслить. Сопроводительные описания написаны для пользователей со специальной подготовкой в области математики.

Предлагаемое читателю пособие дает основы знаний, необходимых для практического использования элементарных методов статистического анализа. Рассматриваются примеры статистического анализа с использованием программных средств.

Выбор конкретных методов статистического анализа и математических методов определяется целью и задачами исследования. Цель большинства математических методов и статистики – установление различия и на его основе проведение классификации, раскрытие взаимосвязи и оценка влияния факторов на состояние или развитие объекта или явления. Итогом таких исследований является изучение тенденций и закономерностей, которые проявляются в массе наблюдений и не могут быть достоверно проанализированы в отдельно взятых случаях.

По виду учетные признаки могут быть качественными или количественными. Качественные, описательные или атрибутивные признаки характеризуют качество отдельных единиц совокупности. Например, категории хозяйств, образование землеустроителей (среднее специальное, высшее) и т. д. Количественные признаки характеризуют числовое выражение в единицах измерения (масса – кг, скорость – км/ч).

Аналитическая оценка взаимосвязи качественных и количественных признаков проводится только после разбиения количественных признаков на качественные группы.

Чрезмерное увеличение объема исходной информации ведет к увеличению «информационного шума» (роста числа помех). Достигнув известного предела, «шум» подавляет искомую информацию.

Механический подход при использовании математических методов недопустим. Каждый из методов анализа имеет свои возможности и ограниченную область применения. Ниже рассмотрим те математические методы, которые используются при решении землеустроительных задач.

Глава 1. ОСНОВЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Статистические материалы могут быть получены при проведении эксперимента, наличии статистической информации об объекте, фондовых материалов, землеустроительных карт, аэрофотоснимков. Чем больший регион занимает объект исследования, тем чаще используются карты, обобщающие материалы, литературные источники.

1.1. ОБРАБОТКА ВАРИАЦИОННОГО РЯДА

Набор цифр в последовательности их увеличения или уменьшения представляет собой *вариационный ряд*. Вариационный ряд может иметь объем от трех и более наблюдений (цифр), которые называются *вариантами*. В зависимости от полноты представления информации по исследуемому объекту вариационный ряд можно характеризовать как *генеральную совокупность*, так и *выборочную совокупность* (выборку).Число ее членов (цифр или вариант) характеризует объем совокупности, который обозначается как *N*, *n.* Множество единиц наблюдения, охватываемое сплошным исследованием во времени или пространстве, называется генеральной совокупностью. Сплошные исследования дорогие и занимают много времени, поэтому проводят выборочные исследования (наблюдения). При выборочном исследовании может быть также достигнута большая глубина и детальность изучения вопроса. Кроме того, при меньшем числе наблюдений уменьшаются вероятности систематических ошибок. Однако выборка должна быть *репрезентативной* (представительной) и *рендомизированной* (методически выдержанной), чтобы объективно охарактеризовать генеральную совокупность. Конечной целью изучения выборочной совокупности является получение информации о генеральной совокупности.

Репрезентативность выборочной совокупности бывает количественной и качественной (структурной). Количественная репрезентативность определяется числом наблюдений, гарантирующих получение статистически достоверных данных. Здесь действует основной постулат закона больших чисел: чем большее число наблюдений, тем больше значений характеристик выборочной совокупности приближаются к аналогичным значениям характеристик генеральной совокупности.

Качественная репрезентативность обозначает структурное соответствие выборочной и генеральной совокупностей. Например, если в составе генеральной совокупности фермерские хозяйства составляют 50 %, то и в выборочной группе их должно быть 50 %.

Для группировки качественных признаков используются альтернативная шкала и шкала рангов (шкала баллов, номиналов, шкала категорий). При статистической сводке все качественные признаки кодируются (1, 2, 3) или им присваивается ранг (1, 2, 3 и т. д.), они приобретают количественное выражение. Ранги, баллы и другие показатели качественных признаков относятся к *непараметрическим*, необходимые величины рассчитываются по специальным формулам. Их нельзя включать в статистическую обработку вместе с количественными характеристиками изучаемого объекта, как и нельзя вычислять средние ранговые величины. Оптимальное число рангов от 6 до 14.

По времени наблюдение может быть текущим (непрерывным) и единовременным (в один и тот же момент времени в разных точках – хозяйства в различных областях).

Отбор объектов для анализа и статистические данные о них можно производить следующими методами: *случайным, направленным (типическим) и смешанным.*

При случайном отборе можно использовать таблицу случайных чисел (прил. 1), в которой представлен набор цифр в четырехзначных столбцах. Двигаясь по столбцу сверху вниз, выписывают по порядку цифры, которые будут соответствовать номеру вариант в статистической совокупности или номеру объекта для проведения исследования.

Направленный отбор заключается в выборе каждой третьей или пятой, или иной варианты из статистической совокупности. Можно таким же способом отбирать объекты для исследования.

Смешанный отбор производят в том случае, когда необходимо выбрать варианты для неоднородного объекта, например, хозяйства большие и малые, разной специализации и т. д.

Иногда сложно определить объем выборки для статистического анализа. В большинстве случаев надежные результаты получают при объеме выборки около 100. Оптимальный объем выборки обычно прямо пропорционален степени увеличения изменчивости признака. Ее объем можно определить по таблице достаточно больших чисел (прил. 2) и расчетным способом. В обоих случаях объем выборки определяется исходя из величины допускаемой вероятности (*P*: 0,95; 0,99 и т. д.), с какой предполагается делать заключение и величины ошибки (точности) опыта (*p*: 1, 2 или 3 % и т. д.).

Расчет объема выборки (*N*) проводят по одной из двух формул:

$$N = \sigma^2 / m^2_M; \quad N = (1{,}96) \cdot V)^2 / p^2, \qquad (1.1)$$

где σ^2 – средний квадрат отклонения (дисперсия); m_M – ошибка среднего арифметического; V – коэффициент вариации, %; p – точность опыта; 1,96 – поправочный коэффициент для уровня вероятности 0,95; 0,99.

Определение объема выборочной совокупности необходимо для достоверной характеристики генеральной совокупности по показателям выборки.

Варианты в выборке практически не совпадают между собой, т. е. они варьируют. Те варианты в начале или конке выборки, которые резко отличаются по величине от соседней варианты, определяются как *артефакт*. Такие варианты исключаются из статистической совокупности и не учитываются при расчете иных показателей. Например, в приведенном вариационном ряду – **2**, 9, 11, 12, 12, 13, 15, **25** – вызывают сомнение варианты 2 в начале ряда и 25 – в конце. Их можно принять за артефакт и исключить при дальнейших расчетах необходимых показателей. Однако такую выбраковку необходимо статистически доказать. В качестве критерия выбраковки может быть использован критерий τ (тау) (прил. 3). Если критерий $\tau_в$ вычисленный (фактический) больше или равен $\tau_т$ табличному ($\tau_в \geq \tau_т$) при объеме выборки N и уровне вероятности P (0,95 или 0,99) или уровне значимости α (0,05 или 0,01), то соответствующие значения варианты (x) выборки следует признать артефактом и исключить из обработки. Значения $\tau_в$ вычисляются по формулам:

$$\tau_1 = (x_2 - x_1) / (x_{n-1} - x_1) - \quad (1.2)$$

для первой варианты в вариационном ряду;

$$\tau_n = (x_n - x_{n-1}) / (x_n - x_2) - \quad (1.3)$$

для последней варианты в вариационном ряду.

В нашем случае $\tau_1 = (9 - 2) / (15 - 2) = 0{,}538$;

$$\tau_8 = (25 - 15) / (15 - 9) = 1{,}666.$$

Сравниваем вычисленные $\tau_в$ с табличными $\tau_т$ по прил. 3. Для объема выборочной совокупности $N = 8$ и α 0,05 табличное значение $\tau_т = 0{,}544$ и для α 0,01 табличное значение $\tau_т = 0{,}683$. Для первой варианты, равной 2, вычисленное $\tau_1 = 0{,}538$ меньше табличных значений $\tau_т$ при обоих значениях α, поэтому она не является артефактом и включается в обработку выборочной совокупности. Для восьмой варианты в ряду, равной 25, вычисленное значение $\tau_в = 1{,}666$ больше, чем табличное $\tau_т$ при обоих значениях α, поэтому варианта 25 признается артефактом и выбраковывается из статистической обработки.

Следующим шагом является вычисление статистических показателей распределения для выборочной совокупности.

1.2. СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ РАСПРЕДЕЛЕНИЯ

Основная задача статистической обработки – вычисление показателей распределения статистической совокупности. Они представлены в общем виде в табл. 1. Их можно разделить на параметрические и непараметрические показатели. К *непараметрическим показателям* относятся мода, медиана, лимит, амплитуда, коэффициент вариации, квантили, частота, частость. Они не включаются в статистическую обработку и служат для приблизительной оценки признака.

Таблица 1.1

Статистические показатели распределения

Показатели	Назначение показателей	Примеры показателей
Средние величины	Описывают положение середины распределения	Центра распределения: Среднее арифметическое Среднее гармоническое Среднее квадратическое Среднее кубическое Среднее геометрическое Средневзвешенное арифметическое Структурные средние: Мода – Мо Медиана – Ме
Показатели разброса	Описывают степень разброса (вариабельности, изменчивости данных)	Лимит – $Lim = x_{max} / x_{min}$ Амплитуда – *Ampl* Дисперсия – σ^2 Среднеквадратическое отклонение – σ Коэффициент вариации – v Квантили

Параметрические показатели представлены случайными величинами с определенными единицами измерения и используются для статистической обработки на основе теории вероятности.

В статистической обработке иногда используют частоту и частость. *Частота* показывает сколько раз встречаются одинаковые значения признака в вариационном ряду. *Частость* – процент вариант от общего числа наблюдений. Она показывает долю частот отдельных вариант от общего числа наблюдений.

Полигон распределения – графический вариационный ряд в системе прямоугольных координат.

Показатели центра распределения (средние)

Средняя величина выражает типичную для данного вариационного ряда величину признака и является равнодействующей всех факторов, влияющих на признак. В ней поглощаются индивидуальные различия вариант в ряду, обусловленные случайными обстоятельствами.

Разница между средними выборок тем больше, чем больше вариабельность признака в статистическом ряду. Рассмотрим показатели центра распределения.

Мода (*Mo*) представляет собой наиболее часто встречающуюся варианту в вариационном ряду. На графике она соответствует максимальной ординате и находится на вершине вариационной кривой. Если вариационный ряд разбит на классы, то мода соответствует максимальной частоте класса, который называется *модальным*. При полимодальном (многовершинном) распределении вариационный ряд имеет несколько значений моды.

Медиана (*Me*) представляет собой среднюю варианту в ранжированном вариационном ряду, которая делит его на две равные части. При нечетном числе вариант середину ряда будет составлять одна варианта (медиана). При четном числе вариант середину ряда образуют две варианты, среднее арифметическое которых будет характеризовать медиану.

При наличии в вариационном ряду сильно отличающихся вариант медиана будет характеризовать середину ряда более точно, чем среднее арифметическое. Мода и медиана используются в тех случаях, когда о выборочных параметрах необходимо иметь ориентировочное представление.

Среднее арифметическое (M, $\bar{o}$) представляет собой величину, сумма положительных и отрицательных отклонений от которой равна нулю. Оно является основной характеристикой статистической совокупности и вычисляется по формуле:

$$M = \sum x_i / N, \qquad (1.4)$$

где $\sum x_i$ – сумма всех вариант совокупности. Среднее арифметическое рассчитывается в тех случаях, когда противопоказано вычислять другие средние.

Пример. Имеем следующие варианты в трех пунктах наблюдений: 10, 15 и 20 ($N = 3$). Среднее арифметическое равно: M = (10 + 15 + 20) / 3 = 15.

Среднее гармоническое ($M_{\text{гар}}$) вычисляется при усреднении меняющихся скоростей процессов (скорость течения воды), показателей обратно пропорциональной зависимости между процессами и явлениями, слож-

ных абсолютных величинах измерений (тонна/километр). Оно рассчитывается по формуле:

$$M_{\text{гар}} = N / \sum (1 / x_i). \qquad (1.5)$$

Его величина меньше, чем величина среднего арифметического. Для вычисления сохраним те же количественные варианты, что и для определения среднего арифметического.

Пример. При перевозке грузов из трех объектов получили следующие величины – 10, 15 и 20 т/км:

$$M_{\text{гар}} = 3 / \sum \left[(1 : 10) + (1 : 15) + (1 : 20)\right] = 13{,}8 \text{ т/км}.$$

Среднее геометрическое ($M_{\text{г}}$) вычисляется в тех случаях, когда в вариационном ряду отдельные значения распределяются в геометрической прогрессии (резко различаются между собой, например, 4 и 16). В данном случае среднее геометрическое равно 8. Оно в два раза меньше 16 и в два раза больше 4. Среднее арифметическое из этих вариант 10, т. е. больше среднего геометрического. При наличии нулевой варианты рассчитывается приближенное среднее арифметическое. Если варианты представлены логарифмами чисел (pH и др.), то вычисляют среднее логарифмическое. Расчет производится по формуле:

$$M_{\text{г}} = \sqrt[n]{x_1 \; x_2 \; \ldots \; x_n}. \qquad (1.6)$$

Пример. Строение стоит 100 тыс. у. е. Одним лицом оно оценивается в 10 млн, другим в 1000 млн. С арифметической точки зрения в первом случае получаем ошибку в 90 млн у. е., во втором – в 900 млн у. е. Если оценивать, во сколько раз ошиблись покупатели, то получаем один ответ в обоих случаях – в 10 раз.

Среднее квадратическое ($M_{\text{кв}}$) используется, когда необходима проверка результатов эксперимента на единство суммарного действия (площадь земельных участков, функциональных зон и т. д.). Вычисляется по формуле:

$$M_{\text{кв}} = \sqrt{\sum x_i^2 \big/ N}. \qquad (1.7)$$

Пример. Имеются данные по величине радиусов трех земельных участков: 10, 15 и 20 м^2. Среднее квадратическое будет равно:

$$M_{\text{кв}} = \sqrt{\sum x_i^2 \big/ N} = \left[\sum (10^2 + 15^2 + 20^2) / 3\right]^{-1/2} = 15{,}56 \text{ м}^2.$$

Среднее кубическое ($M_{\text{куб}}$) применяется при проверке на единство суммарного действия, например, при нахождении средней величины объема. Вычисляется по формуле:

$$M_{\text{куб}} = \sqrt[3]{\sum x_i^{\,3} \big/ N}. \qquad (1.8)$$

Пример. Кубатура строений по трем участкам составляет 10, 15, и 20 м^3. Определяем среднее кубическое по формуле:

$$M_{\text{куб}} = \sqrt[3]{\sum x_i^3 / N} = \sqrt[3]{\sum (10^3 + 15^3 + 20^3) / 3} = 16{,}03 \text{ м}^3.$$

Величина средней кубической максимальна по сравнению с другими средними и находится в ряду справа от всех средних: $M_{\text{гар}} < M_{\text{г}} < M < M_{\text{кв}} < M_{\text{куб}}$.

Средневзвешенная ($M_{\text{взв}}$). Сгруппированный вариационный ряд по классам иногда называют взвешенным из-за той роли, которую выполняют частоты. Чем больше частота вариант в классе, тем больший вес она имеет в характере распределения числового ряда. Среднее арифметическое, рассчитанное в этом ряду, называют взвешенным средним:

$$M_{\text{взв}} = \sum \left[(x_1 \cdot f_1) + (x_2 \cdot f_2) + \ldots + (x_n \cdot f_n) \right] / \sum f_i, \qquad (1.9)$$

где x_n – варианты; f_i – частоты по классам.

Если совокупность вариант разбита на несколько неравных по численности групп, то среднюю арифметическую вычисляют для каждой группы. Затем их объединяют, определяя *общее среднее* ($M_{\text{общ}}$):

$$M_{\text{общ}} = \sum M_j \cdot n_j / \sum n_j, \qquad (1.10)$$

где M_j – среднее по группам; n_j – число вариант в группе.

Среднее логарифмическое, как и среднее арифметическое, вычисляется в случае представления вариант в виде логарифмических чисел (pH и др.).

Вычисление ошибки среднего приведено в п. 1.4.

Показатели рассеивания вариант

Для характеристики распределения в вариационном ряду недостаточно лишь средней величины. В разных по величине вариантах двух выборок среднее может быть равнозначной величиной:

$$-100; -20; 100; 20; \; M = 0,$$

$$0{,}1; -0{,}2; 0{,}1; \; M = 0.$$

Для получения более полного представления о выборочных совокупностях используют показатели рассеяния вариант, или разнообразия признаков: лимит, размах варьирования (амплитуда), среднеквадратическое (стандартное) отклонение, средний квадрат отклонений (дисперсия), коэффициент вариации, квантили. Эти показатели признаков характеризуют различную степень и особенности разброса.

Лимит указывает границы вариационного ряда:

$$Lim = x_{\max} \div x_{\min}.$$

Амплитуда (вариационный размах, размах варьирования) – разность между максимальным и минимальным значениями вариант:

$$Ampl = x_{\max} - x_{\min}.$$

Чем ближе минимальные и максимальные варианты к среднему и чем меньше амплитуда, тем меньше степень разнообразия между переменными в вариационном ряду, тем надежнее характеризуют статистические показатели искомую закономерность.

Более точно степень разнообразия признака следует характеризовать другими показателями. Среднеквадратическое отклонение и дисперсию используют как составляющие параметры нормального распределения при вычислении ряда параметрических статистических критериев.

Среднеквадратическое отклонение, или *сигма* (σ) показывает степень рассеяния значений статистической совокупности около среднего значения, а точнее, интервал ($M \pm \sigma$), в который входит до 75 % вариант выборочной совокупности. Считается, если 75 % вариант выборки находится в пределах $M \pm \sigma$, то это соответствует норме (стандартному отклонению); если в пределах $M \pm 2\sigma$, то имеется незначительное отклонение от нормы; если выходит за пределы $M \pm 3\sigma$, то можно утверждать о наличии анормального явления, процесса. Величина сигмы зависит прямо пропорционально от разброса вариант в вариационном ряду. Чем больший разброс, тем больше значение сигмы. Однако он не дает оценки разброса, как и дисперсия: велик, средний или мал.

Среднеквадратическое отклонение используется для:

– оценки вариант одноименных выборок при близких средних: чем больше сигма, тем больший разброс вариант в совокупности, соответственно среднее арифметическое менее типично для данного вариационного ряда;

– для оценки типичности средней величины в ряду, используя правило трех сигм ($M \pm 3\sigma$);

– для определения доверительных интервалов статистических коэффициентов и репрезентативности выборочных исследований.

Недостаток сигмы, как и дисперсии, в том, что они представляют собой абсолютную именованную величину, поэтому их нельзя использовать при сравнении выборочных совокупностей, выраженных в различных единицах измерения. Для этой цели подходит коэффициент вариации.

Среднеквадратическое отклонение можно определить двумя путями:

$$\sigma = \sqrt{\sum (x_i - M_x)^2 / (N-1)}, \qquad (1.11)$$

$$\sigma = (x_{\max} - x_{\min}) / 6, \qquad (1.12)$$

где $(x_i - M_x)$ – отклонение от среднего индивидуальных вариант; N – объем выборочной совокупности.

Формулу (1.12) можно использовать для приближенного расчета сигмы. Алгебраически сигма представляет собой корень квадратный из дисперсии.

Пример. Получены следующие данные по площади полей севооборота: 20, 20, 22, 23, 24, 25, 25, 26, 27, 28, 30 (м^2).

Для расчета сигмы составляем таблицу 1.2 исходных данных. Подставив в формулу (1.11) данные, определяем сигму:

$$\sigma = \sqrt{100{,}85/10} = 3{,}17 \text{ м}^2.$$

Таблица 1.2

Форма записи и расчета среднеквадратического отклонения

x_i	$x_i - M_x$	$(x_i - M_x)^2$	x_i	$x_i - M_x$	$(x_i - M_x)^2$
20	–4,73	22,37	26	1,27	2,54
20	–4,73	22,37	27	2,27	5,15
22	–2,73	7,45	28	3,27	10,69
23	–1,73	2,99	30	5,27	27,77
24	–0,73	0,53			
25	0,27	0,07	$\sum x_i = 270$	$\sum$ –2,23	$\sum(x_i - M_x)^2 = 102$
25	0,27	0,07	$M_{кв} = 24{,}73$		

Среднее квадратическое равно 24,73 м. Если значение сигмы 3,17 прибавить к среднему арифметическому и вычесть из него, то определим граничные значения, в которых будет находиться определенная часть вариант (до 75 %) исследуемой статистической выборки (24,73 ± 3,17). В этот интервал (от 21,56 до 27,90) вошли варианты 22, 23, 24, 25, 25, 26, 27. Это означает, что 68 % вариант в выборке находится в пределах от 21,56 (24,73 – 3,17) до 27,90 м (24,73 + 3,17 = 27,90). Лишь 32 % вариант выходит за указанные пределы.

Средний квадрат отклонений, или *дисперсия* указывает колебание значений признака внутри выборочной совокупности через отклонение всех вариант от среднего значения, т. е. показывает интервал, в который входят все варианты выборки (100 %). Для получения обобщающей характеристики числового ряда нельзя использовать сумму отклонений от среднего, так как сумма положительных и отрицательных отклонений равна или близкая к нулю.

Этого можно избежать, если отклонения от среднего возвести в квадрат. Возведенные в квадрат положительные и отрицательные отклонения дают положительные значения при усреднении всех квадратов отклонений вариационного ряда, и получается *средний квадрат отклонений*, который называют *дисперсией*.

Поэтому все отклонения от среднего возводятся в квадрат и суммируются: $\sum(x_i - M_x)^2$. При усреднении всех отклонений числового ряда пу-

тем деления на (N – 1) получаем средний квадрат отклонений, или дисперсию (D, σ^2).

Если вычислена сигма (σ), то дисперсию получаем путем возведения ее в квадрат: σ^2.

При упрощенном способе расчета дисперсии не вычисляют отклонений вариант от среднего ($x_i - M_x$), используя следующий расчет:

$$\sigma^2 = \sum x_i^2 / N - M^2,$$

где $\sum x_i^2$ – сумма квадратов всех вариант выборки; M^2 – квадрат среднего арифметического; N – число вариант в выборке.

Более точно значение дисперсии вычисляется с использованием данных в табл. 1.2 по формуле:

$$\sigma^2 = \sum (x_i - M_x)^2 / (N - 1). \quad (1.13)$$

Недостатком дисперсии является несоответствие ее размерности (квадраты отклонений) и размерности единиц измерения вариант выборки. Если варианты выражены в килограммах, то дисперсия дает квадрат этой меры. Дисперсию можно заменить на среднеквадратическое отклонение (δ), или стандартное отклонение, или стандарт распределения. Форма записи исходных данных для вычисления дисперсии такая же, как и для сигмы (см. табл. 1.2). Подставив значения из таблицы в формулу, получим значение дисперсии:

$$\sigma^2 = 102 / 10 = 10{,}2 \text{ м}^2.$$

Исходя из величины дисперсии, можно определить интервал, в пределы которого входят все варианты выборки: $M \pm \sigma^2$, от 14,73 м (24,73 – 10,0) до 34,73 м (24,73 + 10,0). В этот интервал вошли 100 % вариант выборочной совокупности.

При объединении нескольких аналогичных выборок в общую выборку можно рассчитать общий средний квадрат отклонений, если имеются сведения о дисперсии по каждой выборке в отдельности:

$$\sigma^2_{\text{общ}} = \sum (N_i - 1) \cdot \sigma^2_i / (\sum N_i - k), \quad (1.14)$$

где σ^2_i – дисперсия индивидуальной выборки; N_i – объем частных выборок; k – число частных выборок.

Пример. Вычислим общий средний квадрат отклонений для четырех выборок, отражающих содержание кальция в озерных водах Беларуси: $\sigma^2_1 = 2$; $N_1 = 8$; $\sigma^2_2 = 2{,}5$; $N_2 = 6$; $\sigma^2_3 = 3{,}0$; $N_3 = 7$; $\sigma^2_4 = 3{,}5$; $N_4 = 8$. По формуле (1.8) имеем:

$$\sigma^2_{\text{общ}} = \frac{(8-1)\cdot 2 + (6-1)\cdot 2{,}5 + (7-1)\cdot 3 + (8-1)\cdot 3{,}5}{(8+6+7+8)-4} = 2{,}76.$$

Если извлечь корень квадратный из полученной величины, получим общее среднее квадратическое отклонение, или сигму ($\sigma_{общ} = 1{,}66$).

Практическое применение дисперсии состоит в следующем:

– для оценки вариабельности рядов распределения;

– для факторного и дисперсионного анализа;

– для статистической оценки двух совокупностей по критерию Фишера.

Дисперсия выражается в тех же единицах, что и варианты выборочной совокупности.

Коэффициент вариации представляет собой относительный показатель разнообразия признаков, выражается в процентах. Он показывает отношение среднеквадратического отклонения к средней арифметической:

$$V = (\sigma / M) \cdot 100. \qquad (1.15)$$

В случаях, когда значение среднеквадратического отклонения не рассчитывается, величина коэффициента вариации может быть определена следующим образом:

$$V = 100 \sqrt{\frac{\sum x_i^2 / (M^2 - N)}{N - 1}}, \qquad (1.16)$$

где $\sum x_i^2$ – сумма квадратов индивидуальных вариант в совокупности.

Чем меньший по размаху варьирования будет признак, тем меньший будет коэффициент вариации для данной совокупности. Соответственно меньшими будут сигма и дисперсия.

Коэффициент вариации позволяет оценить вариабельность (разброс) признака в нормированных границах. Если его значение меньше 10 %, то разброс вариант относительно средней арифметической считается слабым, при 10–30 – средним, 30–60 – высоким, 60–100 – очень высоким, более 100 % – аномальным.

О преимуществе использования коэффициента вариации при оценке разнородных признаков можно судить по табл. 1.3.

Таблица 1.3

Сравнительная оценка состава работников предприятия

Учетный признак	Среднее арифметическое, *M*	Среднеквадратическое отклонение, σ	Коэффициент вариации, *V*
Стаж работы (лет)	8,7	2,8	32,1
Возраст (лет)	37,2	4,1	11,0
Образование (класс)	9,2	1,1	11,9

В табл. 1.3 абсолютные величины средних и сигмы близки по стажу работы и образованию. Однако по коэффициенту вариации сходны по возрасту и образованию. В данном случае сравнение по сигме проводить не корректно, так как все три признака разнородны и не сравнимы между собой. Выручает неименованный коэффициент вариации, который позволяет оценить разброс признака в нормированных границах.

Коэффициент вариации нельзя рассчитывать по формулам 1.15, 1.16 при наличии вариант признака с отрицательным числом (чередование положительных и отрицательных температур, отметка поверхности ниже уровня воды в океане и др.). В таких случаях коэффициент вариации рекомендуется вычислять по формуле с учетом модуля:

$$V = 100\ \sigma / |x_i| + M, \qquad (1.17)$$

где $|x_i|$ – модуль наименьшей отрицательной величины без учета знака.

В данном случае имеется в виду, что при вычислении коэффициента вариации среднее арифметическое и среднеквадратическое отклонения должны быть представлены в виде отрезков на числовой оси. Приведем алгоритм вычисления коэффициента вариации для величин с разными знаками.

Пример. Температура воздуха в течение суток в октябре составила (в градусах Цельсия): –4, –3, –1, +1, +3. Среднее арифметическое равно –0,6, среднеквадратическое отклонение 1,95. Если не учитывать наличия интервальной шкалы и определять коэффициент вариации по формуле (1.9), то получим следующую величину:

$$V = (1{,}95 \cdot 100) / (-0{,}6) = -325\ \%.$$

Результаты противоречат исходным данным, которые фактически характеризуются небольшим размахом варьирования температур в течение суток. Если среднее арифметическое представить как отрезок от точки –4 до –0,6, то оно будет равно: |–4| + (–0,6) = 3,4. Используя формулу (1.17), получаем коэффициент вариации, соответствующий условиям задачи:

$$V = (100 \cdot 1{,}95) / \big(|-4| + (-0{,}6)\big) = 54{,}16\ \%.$$

Квантили. В открытых вариационных рядах и рядах распределения качественных признаков для сжатого описания распределений используется другой параметр разброса – *квантиль* (синонимы: перцетиль, персентиль). Этот параметр может использоваться для перевода количественных признаков в качественные. В практике статистического анализа наиболее часто используются следующие квантили:

$V_{0,5}$ – медиана;

$V_{0,25}$, $V_{0,75}$ – квантили четверти, соответственно нижняя и верхняя квантиль;

$V_{0,1}$, $V_{0,2}$, … $V_{0,9}$ – децили (десятые);

$V_{0,01}$, $V_{0,02}$, … $V_{0,99}$ – процентили, или центили (сотые).

Квантили делят область возможных изменений вариант в выборке на определенные интервалы. Статистическая суть квантилей лучше раскрывается при построении графика.

При оценке явлений с помощью квантилей придерживаются следующих разграничений:

● Показатели меньше 3-го центиля оцениваются как резко пониженные (недостающие), между 3-м и 10-м центилем – как пониженные, между 10-м и 25-м центилем – ниже среднего, между 25-м и 75-м – средние, между 75-м и 90-м – выше среднего, между 90-м и 97-м – повышенные, выше 97-го – резко повышенные.

● Числовые значения признака в квантилях можно представить в виде диаграммы. С помощью пакета анализа MsExcel более точно рассчитывают квартили и другие виды квантилей.

Проверка статистических гипотез

Методологической основой исследования является формулировка рабочей гипотезы, которую в ходе работы можно было бы признать истинной или отвергнуть (признать ложной).

Статистической называют гипотезу о виде неизвестного распределения или о параметрах распределений, например:

● генеральная совокупность распределена по закону Пуассона;

● средние арифметические двух совокупностей не равны между собой;

● дисперсии (разброс вариант) двух совокупностей равны между собой и др.

Выдвинутую гипотезу называют *основной* или *нулевой* (H_0). Гипотезу, которая противоречит нулевой и является ее логическим противоречием, называют *конкурирующей* или *альтернативной* (H_1).

Короткая запись простой гипотезы следующая:

$$H_0 : M = 1; H_1 : M \neq 15$$

В качестве нулевой гипотезы обычно принимают гипотезу об отсутствии различий.

Сложная гипотеза записывается так:

$$H : D > 15; H : D > 17; H : D > 19 \text{ и т. д.}$$

Выдвинутая гипотеза проверяется статистическими методами. При проверке могут быть допущены ошибки двух родов: 1) отвергается правильная гипотеза; 2) принимается неправильная гипотеза.

Вероятность совершить ошибку первого рода называют *уровнем значимости* (α), который принимается не выше 0,05 (5 %). Это означает, что в 5 случаях (5 %) из 100 мы рискуем допустить ошибку первого рода.

Значимость ошибки второго рода обозначают символом β.

Для проверки нулевых гипотез используют *статистические критерии* (*К*). Величины фактических критериев ($К_ф$) сравнивают со значениями табличных ($К_{табл}$). В роли критерия может выступать критерий Фишера (отношение дисперсий), критерий Стьюдента (*t*), критерий соответствия (χ^2 – хи-квадрат), с учетом допустимой области его применения, которая задается законом его распределения. Критерии могут быть параметрическими и непараметрическими. Непараметрические критерии не подчиняются закону нормального распределения.

Доверительная значимость. Статистическая значимость выборочных характеристик представляет собой меру уверенности в их «истинности». Уровень значимости находится в убывающей зависимости от надежности результата. Более высокая статистическая значимость ($\alpha > 0,05$) соответствует более низкому уровню доверия к найденной в выборке характеристике. Именно уровень значимости представляет собой вероятность ошибки, связанной с распространением наблюдаемого результата на всю генеральную совокупность. Выбор порога значимости, выше которого результаты отвергаются как статистически не подтвержденные, во многом произвольный. Это зависит от традиций и накопленного практического опыта. Верхняя граница $\alpha < 0,05$ статистической зависимости содержит большую вероятность ошибки (5 %).

Уровень значимости выражает вероятность нулевой гипотезы, т. е. вероятность того, что выборочные и генеральные средние не отличаются друг от друга. Чем выше уровень значимости, тем меньше можно доверять утверждению, что различия существуют (табл. 1.4).

Таблица 1.4

Интерпретация уровней значимости (*α*)

Показатели значимости (α)	Интерпретация
$\geq 0,1$	Данные согласуются с нулевой гипотезой (H_0)
$\geq 0,05$	Есть сомнения в истинности как нулевой (H_0), так и альтернативной гипотез (H_1)
$< 0,05$	Нулевая гипотеза (H_0) может быть отвергнута
$\leq 0,01$	Нулевая гипотеза (H_0) может быть отвергнута. Сильный довод
$\leq 0,001$	Нулевая гипотеза (H_0) почти наверняка не подтверждается. Очень сильный довод

Теоретические распределения. Важнейшее звено статистического анализа – аналитическое описание кривой конкретного распределения в виде закона распределения.

Теоретическим называют распределение, которое выбирается как образец (стандарт) для описания закона фактического распределения.

Встречаются дискретные (прерывные) и непрерывные распределения. Наиболее распространенные среди дискретных распределений – биноминальное и пуассоновское распределения. Среди непрерывных – нормальное и связанные с ним распределения Стьюдента, хи-квадрат χ^2 и *F*-распределение Фишера.

Биноминальное распределение (распределение Бернулли) возникает, когда оценивается, сколько раз происходит некоторое событие в серии определенного числа независимых, выполняемых в одинаковых условиях наблюдений.

Распределение Пуассона появляется в ситуациях, когда в течение определенного отрезка времени или на определенном пространстве происходит случайное число каких-либо событий (число радиоактивных распадов, выпадение частиц аэрозоля). Основное отличие этого закона распределения – резко выраженная асимметрия. Если в ходе статистической обработки дискретных рядов получены среднее и дисперсия, которые равны между собой, то распределение можно считать подчиняющимся закону Пуассона. Биноминальное и Пуассона распределения сходны с нормальным распределением.

Нормальное распределение (Гаусса) используется для приближенного описания явлений, которые носят вероятностный, случайный характер.

Параметры M и δ описывают различные модификации нормального распределения. Статистические критерии, которые используют эти параметры, принято называть параметрическими критериями.

При нормальном распределении в области $\geq M \pm \delta$ оказывается 68,3 % всех вариант, в пределах $\pm 2\delta$ – 95,5 % вариант, в пределах $\pm 3\delta$ – 99,7 %. Эта закономерность описывается как правило тремя сигмами.

1.3. ОЦЕНКА СТАТИСТИЧЕСКИХ ПАРАМЕТРОВ

Оценка в статистике позволяет установить степень соответствия параметров выборки аналогичным параметрам генеральной совокупности, а также дать оценку точности опыта при проведении исследований.

Ошибка параметра. Степень соответствия параметров выборки параметрам генеральной совокупности оценивается *ошибкой* (m) того или

иного параметра выборочной совокупности. Например, по мере увеличения числа наблюдений средние выборок и другие параметры все более приближаются к этим параметрам генеральной совокупности.

Ошибка записывается вместе с оцениваемым параметром, например, $M \pm m_M$, $\sigma \pm m_\sigma$, $r \pm m_r$. Это значит, что она указывает на интервал, в котором находится этот параметр генеральной совокупности. Значит, чем меньше ошибка, тем больше соответствует величина параметра выборки величине этого параметра генеральной совокупности. Нулевая ошибка указывает на совпадение величин параметра выборки и этого параметра генеральной совокупности. Расчет ошибок параметров проводится по различным формулам. Приведем расчеты ошибок важнейших статистических параметров.

Стандартная ошибка средней арифметической:

$$m_M = \sqrt{\frac{\sum (x_i - M_x)^2}{N(N-1)}}, \text{ или } m_M = \sqrt{\frac{\sigma^2}{N}}. \qquad (1.18)$$

Ошибка среднеквадратического отклонения определяется:

$$m_\sigma = \sigma / \sqrt{2(N-1)}. \qquad (1.19)$$

Ошибка дисперсии вычисляется путем возведения в квадрат ошибки среднеквадратической.

Ошибка коэффициента корреляции рассчитывается:

$$m_r = \frac{V}{\sqrt{N}} \sqrt{1/2 + (V/100)^2}. \qquad (1.20)$$

Поскольку параметр m характеризует ошибку утверждения (прогноза) о том, что выборочное среднее равно среднему генеральной совокупности, то чем выше требование к вероятности этого вывода, тем шире должен быть обеспечивающий точность такого прогноза интервал, называемый *доверительным интервалом*. Его величина задается вероятностью безошибочного прогноза, которую принято называть *доверительной вероятностью* (*уровнем вероятности,* надежностью опыта, вероятностью безошибочного прогноза). В научных исследованиях допускается доверительная вероятность (P) не менее 95 % (0,95 частей от 1). В этих случаях P для средних арифметических при достаточно большом числе наблюдений ($N > 30$) равен $\pm 2m$. Предельная ошибка выборки $\Delta = M \pm 2m$. При доверительной вероятности 99 % (0,99) доверительный интервал составит $\pm 3m$, $\Delta = M \pm 3m$. По иному в отношении доверительного интервала можно сказать так: *он показывает, какой процент вариант выборки подтверждает искомую статистическую закономерность.*

Каждому значению доверительной вероятности соответствует свой *уровень значимости* (α). Он выражает вероятность нулевой гипотезы: вероятность того, что выборочная и генеральная средние не отличаются друг от друга. Иначе говоря, чем выше уровень значимости, тем меньше можно доверять утверждению, что различия существуют. Следовательно, *он показывает, какой процент вариант выборки отвергает искомую статистическую закономерность*. Уровень значимости дополняет доверительную вероятность, например, α = 5 % (0,05) дополняет уровень вероятности 95 % (0,95). В сумме они составляют 100 % (1).

В таблицах приложения приводятся численные значения для P или α соответственно: 0,95 и 0,99; 0,05 и 0,01. В этих случаях при интерпретации мы можем утверждать нулевую гипотезу (H_0).

Оценка точности опыта. При исследовании методического характера необходимо давать им оценку по показателю *точности опыта* (p). Его смысл состоит в установлении величины ошибки среднего арифметического (m_M) в процентах от величины среднего арифметического (M). Показатель точности опыта можно определить по одной из двух формул:

$$p = (m_M / M) \cdot 100; \; p = V / \sqrt{N}, \qquad (1.21)$$

где V – коэффициент вариации.

Опыт считается достаточно точным, если $p < 3$ %, удовлетворительным – при его величине 3–5 %. При величине точности опыта более 5 %, к полученным выводам следует относиться осторожно и увеличить число повторностей в опыте или его повторить. В полевых опытах с растениями, результаты которых обрабатываются с использованием метода дисперсионного анализа. Точность опыта допускается до 3 %. Некоторые приборы при анализе химического состава вещества могут давать погрешность до 15 %, например, в полуколичественном спектральном анализе.

Ошибка точности опыта вычисляется следующим образом:

$$m_p = \pm p \sqrt{(1/2N) + (p/100)^2}. \qquad (1.21)$$

Пример. Среднее арифметическое из четырех повторностей фитомассы многолетних трав в полевых опытах составило M = 350 ц/га, ошибка среднего арифметического m_M = 5 ц/га, N = 16. Используя формулу (1.21), определим точность опыта:

$$p = (m_M / M) \cdot 100 = (5 / 350) \cdot 100 = 1{,}42\ \%.$$

Полученная величина точности опыта 1,42 % достаточно точная. Опыт считается положительным.

1.4. СТАТИСТИЧЕСКИЕ КРИТЕРИИ УСТАНОВЛЕНИЯ РАЗЛИЧИЙ

Проведение землеустроительных исследований предполагает установление сходства или различия между одноименными генеральными совокупностями изучаемых систем по выборочным совокупностям. Сопряженный анализ одноименных признаков в выборках используется для классификации и районирования по одному или нескольким параметрам. Возникает необходимость применения объективного метода выделения классификационных групп или районов на основе методов математической статистики с использованием критериев достоверности. Если достоверность различия между выборочными совокупностями доказана, то генеральные совокупности, сравниваемые по какому-либо признаку, выделяют как самостоятельные. В случае отсутствия достоверных различий их объединяют в одну группу.

Различие между двумя выборками устанавливается с помощью ряда критериев: t – распределение Стьюдента, наименьшего существенного различия (НСР), F-распределения Фишера, критерия соответствия (χ^2).

Каждый из критериев применяется при определенных условиях, которые задаются целью исследования. Несоблюдение указанных условий может привести к ошибочным выводам.

Прежде, чем приступать к статистической обработке и расчету критериев различия, следует убедиться в отсутствии артефакта в сравниваемых выборках. Если в малых совокупностях распределение нормально, то для установления артефакта достаточно использовать правило трех сигм. Согласно этому правилу, в пределах $M \pm 3\sigma$ находится 99,7 % всех вариант выборки. Если крайние варианты попадают в этот интервал, то они включаются в статистическую выборку, так как не являются артефактом. Наличие артефакта можно проверить также по формулам (1.2, 1.3).

Критерий Стьюдента. Используется для оценки сходства или различия между выборочными совокупностями по разности величин их средних ($d = M_{\text{большая}} - M_{\text{меньшая}}$) и ее отношения к ошибке этой разности (m_d) при условии распределения вариант в группах по закону нормального или логнормального распределения, подтверждается равенство разброса вариант в выборке (близкие дисперсии сравниваемых выборок). Не допускается применения критерия в случае балльного характера сравниваемых числовых признаков.

Конкретная методика оценки для установления различий по критерию Стьюдента зависит от вида выборочных совокупностей: *независимых (несвязанных)* или *сопряженных (парных)* выборок, а также для ус-

тановления различия между выборочными и генеральными (теоретическими стандартами) средними.

Независимые статистические совокупности могут быть получены на одном или нескольких объектах, но *при одинаковых условиях* проведения эксперимента: например, сравнение экономического показателя в хозяйстве или на предприятии за интересуемый период между собой; сравнение чистого дохода в хозяйствах с одинаковым экономическим развитием, но расположенных на значительном расстоянии. При сравнении независимых выборочных совокупностей объемы выборок могут быть одинаковы ($N_1 = N_2$) или разные ($N_1 \neq N_2$). В двух сравниваемых независимых выборках с одинаковым или разным объемом наблюдений *степень свободы* определяется по формуле:

$$\nu = (N_1 - 1) + (N_2 - 1) = N_1 + N_2 - 2. \quad (1.22)$$

При *малых объемах независимых совокупностей, если дисперсии сравниваемых выборок нельзя считать одинаковыми*, число степеней свободы определяется сложнее:

$$\nu = \frac{1}{u^2/(N_{x1}-1) + (1-u)^2/(N_{x2}-1)}, \quad (1.23)$$

где $u = m_{x1}^2/(m_{x1}^2 + m_{x2}^2)$; m_{x1} и m_{x2} – ошибки среднего первой и второй выборок соответственно.

Сопряженные статистические совокупности получают на одном или на разных объектах, но в разных условиях. Например, сравнение прибыли фермерских и подсобных хозяйств в любом районе или фермерских хозяйств Витебской и Гомельской области. Объем сравниваемых выборок должен быть одинаков ($N_1 = N_2$). Определение *степени свободы* для *сопряженных* выборок определяется как:

$$\nu = N_{\text{пар}} - 1. \quad (1.24)$$

Ошибка разности между средними выборок (m_d) в зависимости от вида (независимые, сопряженные) и объема наблюдений рассчитывается по разным формулам. Рассмотрим их ниже.

Вариант первый. Сравниваемые выборки имеют одинаковый объем наблюдений ($N1 = N2$) и независимы:

$$m_d = \sqrt{m_{x1}^2 + m_{x2}^2}, \quad (1.25)$$

где m_{x1} и m_{x2} – ошибка средней арифметической первой и второй выборки.

Критерий Стьюдента определяют по формуле:

$$t = d / m_d = (M_{\text{большая}} - M_{\text{меньшая}}) / m_d. \quad (1.26)$$

Сопоставляя критерий Стьюдента вычисленный с табличным устанавливают или отвергают с некоторой долей уверенности различия между средними арифметическими выборок.

Пример. При исследовании глубины расчленения рельефа в северной (x_1) и центральной (x_2) провинциях Беларуси необходимо установить, объединять их в один район по землеустроительным условиям или различать их как самостоятельные. Исходные данные и их обработка приводятся в табл. 1.5. Из полученной информации по средним арифметическим (M_{x1} = 16,6 и M_{x2} = 15,2 м) различие по глубине расчленения рельефа можно признать как существенным, так и несущественным. Для объективных выводов используем критерий Стьюдента.

Таблица 1.5

Форма обработки вариант в независимых совокупностях

X_{i1}	$X_{i1} - M_{x1}$	$(X_{i1} - M_{x1})^2$	X_{i2}	$X_{i2} - M_{x2}$	$(X_{i2} - M_{x2})^2$
20	3,4	11,56	17	1,8	3,24
17	0,4	0,16	16	0,8	0,64
16	–0,6	0,36	15	–0,2	0,04
15	–1,6	2,56	14	–1,2	1,44
15	–1,6	2,56	14	–1,2	1,44
∑ 83	0	17,20	76	0	∑ 6,80
M_{x1} = 16,6			M_{x2} = 15,2		

Определяем разницу между средними: d = 16,6 – 15,2 = 1,4. Ошибки по каждой выборке равны:

$$m_{x1} = \sqrt{\sum (x_{i1} - M_{x1})^2 \big/ N_{x1}(N_{x1} - 1)} = \sqrt{\sum (17,2) \big/ 5(5-1)} = 0,93;$$

$$m_{x2} = \sqrt{\sum (x_{i2} - M_{x2})^2 \big/ N_{x2}(N_{x2} - 1)} = \sqrt{6,8/20} = 0,58.$$

Ошибка разности средних составляет:

$$m_d = \sqrt{m_{x1}^2 + m_{x2}^2} = \sqrt{0,93^2 + 0,58^2} = 1,20.$$

Полученные данные подставляем в формулу (1.26) и вычисляем $t_ф$ = 1,4 / 1,2 = 1,17. Число степеней свободы $\nu = N_{x1} + N_{x2} - 2 = 5 + 5 - 2 = 8$.

Сопоставляем табличные значения критерия Стьюдента 2,31 и 3,36 (см. прил. 4) при P = 0,95 и 0,99 для степени свободы ν = 8 с фактическим (расчетным) $t_ф$ = 1,17. Поскольку t_m (2,31 и 3,36) > $t_ф$ (1,17) при обоих уровнях значимости, то разность между средними признается недостоверной (несущественной). Следовательно, при выделении землеустроительных районов по глубине расчленения рельефа оба района объединяем в одну группу.

Вариант второй. Сравниваемые *независимые* совокупности имеют различие по объему ($N_1 \neq N_2$). Порядок вычисления критерия Стьюдента такой же, как и при установлении достоверности в независимых выборках с одинаковым числом наблюдений. Различие состоит в вычислении по другой формуле ошибки разности средних:

$$m_d = \sqrt{\frac{\left(\sum (x_{i1} - M_{x1})^2 + \sum (x_{i2} - M_{x2})^2\right) \cdot (N_{x1} + N_{x2})}{(N_{x1} + N_{x2} - 2) \cdot (N_{x1} \cdot N_{x2})}}. \qquad (1.27)$$

Вариант третий. Сравниваемые *сопряженные* совокупности имеют одинаковый объем выборки ($N_1 = N_2$). Ошибка разности средних определяется по формуле:

$$m_d = \sqrt{\frac{\sum (d_i - d)^2}{N_{\text{пар}}(N_{\text{пар}} - 1)}}. \qquad (1.28)$$

Обозначения для формул (1.27) и (1.28): x_{i1} и x_{i2} – индивидуальные значения вариант первой и второй выборок соответственно; M_{x1} и M_{x2} – средние первой и второй выборочной совокупности соответственно; N_{x1} и N_{x2} – объем выборки первой и второй соответственно; d_i – разность между индивидуальными сопряженными вариантами в выборках; d – разность между средними сопряженных выборок.

Пример для сопряженных наблюдений. Сравним глубину расчленения рельефа земельных участков в пределах конечно-моренного (x_1) и донно-моренного (x_2) ландшафта. Для обработки данных составляем исходную табл. 1.6.

Число пар в выборках $N_n = 5$. Разность между средними арифметическими сопряженных выборок $d = 16{,}6 - 15{,}2 = 1{,}4$. Ошибку разницы средних рассчитываем по одной из формул:

$$\sqrt{\sum (d_i - d)^2 \big/ N_n(N_n - 1)} = \sqrt{3{,}2/5(5-1)} = 0{,}40;$$

$$m_d = \sqrt{\frac{\sum d_i^2 - (\sum d_i)^2 / N_n}{N_n(N_n - 1)}} = \sqrt{\frac{13 - 7^2/5}{5(5-1)}} = 0{,}40.$$

Результаты расчетов по приведенным формулам не выявили расхождений. Критерий Стьюдента получим следующий: t = 1,4 / 0,40 = 3,5. Число степеней свободы $\nu = N_{\text{п}} - 2 = 5 - 2 = 3$. Для $\nu = 3$ при P 0,95 и 0,99 табличное значение критерия Стьюдента 3,18 и 5,84 соответственно (см. прил. 4). Поскольку $t_{\text{ф}} > t_m$ при $P_{0,95}$, то различие по глубине расчленения рельефа в сравниваемых ландшафтах признается существенным. Такие ландшафты образуют самостоятельные группы.

Если при проведении опыта не учитывать *сопряженность и независимость* выборок, то можно получить противоположный вывод.

При сравнении средних, полученных на основе большого объема наблюдений при соблюдении нормального распределения, определение достоверности различий средних можно выполнить упрощенно:

$$(M_1 - M_2)^2 / (m_1^2 + m_2^2) \geq 9. \quad (1.29)$$

Таблица 1.6

Форма обработки данных сопряженных наблюдений

X_{i1}	X_{i2}	d_i	d_i^2	$d_i - d$	$(d_i - d)^2$
20	17	3	9	+1,6	2,56
17	16	1	1	−0,4	0,16
16	15	1	1	−0,4	0,16
15	14	1	1	−0,4	0,16
15	14	1	1	0,4	0,16
∑83	∑76	∑7	∑13	∑0	∑3,20
$M_{x1} = 16{,}6$	$M_{x2} = 15{,}2$				
$d = 1{,}4$					

Различия средних арифметических можно считать статистически достоверными, если получена величина 9 и более, если меньше – недостоверными.

Наименьшая существенная разность (НСР). Используется в дисперсионном анализе. Она показывает то минимальное различие между средними, начиная с которого при выбранном уровне вероятности сравниваемые средние существенно отличаются друг от друга. Величина критерия выражается в тех же единицах, что и сравниваемые средние выборочных совокупностей и определяется по формуле:

$$\text{НСР} = t_{\text{табл}} \cdot m_d, \quad (1.30)$$

где m_d – ошибка разницы средних; $t_{\text{табл}}$ – табличное значение критерия Стьюдента при уровне вероятности 0,95 или 0,99 и степени свободы, определяемой экспериментом.

Если разность между сравниваемыми средними в условиях эксперимента больше или равна величине НСР при P 0,95 или 0,99, то различие существенно. Используя предыдущий пример по глубине расчленения рельефа, проверим достоверность разницы между средними арифметическими с использованием критерия НСР для случаев независимого и сопряженного наблюдений по формуле (1.30):

$\text{НСР}_{0,95} = 2{,}31 \times 1{,}20 = 2{,}77$ м; $\text{НСР}_{0,99} = 3{,}36 \times 1{,}20 = 4{,}04$ м (для независимых наблюдений);

$НСР_{0,95} = 3,18 \times 0,40 = 1,27$ м; $НСР_{0,99} = 5,84 \times 0,40 = 2,33$ м (для сопряженных наблюдений).

Разница между средними арифметическими глубины расчленения рельефа при независимых и сопряженных наблюдениях одна и та же (1,4 м). Сравнивая ее с величиной НСР, приходим к тем же выводам, что и при использовании критерия Стьюдента.

Критерий Фишера. В выборочных совокупностях дисперсии могут существенно отличаться друг от друга. В таких случаях установление различий между выборочными совокупностями проводится по критерию Фишера (*F* – положительное асимметричное распределение). Расчет производится по формуле:

$$F = \sigma^2_{\text{большая}} / \sigma^2_{\text{меньшая}}. \quad (1.31)$$

Если величина расчетного критерия Фишера ($F_ф$) не превышает величины приведенного в таблице (F_m) (прил. 5), то различие между сравниваемыми дисперсиями считается недостоверным. При $F_ф > F_m$ эти дисперсии достоверно различны, как и сравниваемые по ним генеральные совокупности. Степень свободы рассчитывается для сравниваемых выборок отдельно по формуле: $\nu = N - 1$.

Пример. Необходимо установить достоверность различия в содержании гумуса в дерново-подзолистой заболоченной суглинистой почве для северной (x_1) и центральной (x_2) провинций Беларуси. Объем выборочных совокупностей одинаков (N_1, N_2). В результате обработки данных получены следующие средние и дисперсии: $M_{x1} = 3,53$ %, $\sigma^2_{x1} = 0,0024$ %; $M_{x2} = 3,32$ %, $\sigma^2_{x2} = 0,00032$ %. Сравниваемые совокупности весьма сходны и можно констатировать отсутствие различия между ними. Однако пределы колебаний вариант в совокупностях существенно различны (более чем в 2 раза). В данном случае для сравнения следует использовать критерий Фишера. В результате вычислительных операций получены следующие результаты: $F_ф = \sigma^2_{x1} / \sigma^2_{x2} = 0,0024 / 0,00032 = 7,5$. Степень свободы одинакова для первой и второй совокупности (5 – 1 = 4). Для *P* 0,95 и 0,99 табличное значение критерия Фишера 6,39 и 15,98 соответственно. Поскольку $F_ф > F_m$, то различие в содержании гумуса по провинциям признается существенным при *P* 0,95.

Критерий Пирсона (хи-квадрат, χ^2). Для оценки соответствия или расхождения полученных эмпирических данных и теоретических (расчетных, прогнозных) распределений применяются статистические *критерии согласия.* Среди них наибольшее распространение получил непараметрический критерий К. Пирсона – хи-квадрат. Его можно использовать с различными формами распределения совокупностей. Как и любой другой статистический критерий, он не доказывает справедливость нуле-

вой гипотезы, а лишь устанавливает с определенной вероятностью ее согласие или несогласие с экспериментальными данными. Критерий применяется при условии наличия не менее 5 наблюдений или частот в каждой группе, классе или совокупности. Малые частоты объединяют. Вычисление проводят по формуле:

$$\chi^2 = \sum (\varphi - \varphi')^2 / \sum \varphi', \qquad (1.32)$$

где φ, φ′ – наблюдения или частоты в опыте соответственно эмпирически или теоретически ожидаемые.

Значения хи-квадрат могут быть только положительными и возрастать от нуля до бесконечности. Если вычисленный критерий хи-квадрат больше табличного (теоретического) значения, нулевая гипотеза, которая предполагает соответствие эмпирического и теоретического распределений, отвергается при $\chi^2_{выч} < \chi^2_{табл}$, нулевая гипотеза принимается.

Упрощенно достоверность различий можно определить по правилу Романовского: нулевая гипотеза отвергается, если неравенство:

$$D = (\chi^2 - \nu) / \sqrt{2\upsilon} > 3. \qquad (1.33)$$

Степень свободы при проверке гипотезы о нормальном распределении вычисляется по формуле $\nu = к - 3$, где $к$ – число классов. Различие между экспериментальными вариантами и теоретическими считается достоверным, если $D > 3$.

Критерий Пирсона тем меньше, чем меньше различаются эмпирические и теоретические частоты. Он не позволяет обнаружить различия, которые скрадывает группировка (объединение малых частот в одну группу). Его удобно использовать, так как не требуется вычисление средних, дисперсий.

Пример. Следует определить число сельских механизаторов с бронхолегочными заболеваниями, обострение болезни у которых связано с условиями работы. Для обработки выборочных вариант составляем табл. 1.5. Всего выявлен 71 больной механизатор из 639 обследованных одного возраста и пола по 9 человек в каждом населенном пункте. Количество обследованных сгруппировано в 9 классов. Поскольку частота в каждом классе φ, φ′ должна быть не менее 5, объединяем первые три и последние два класса в столбцах 2 и 3. Получаем новые классы с частотами 11 и 13 (всего по 6 классов распределения). Частоты в новых классах выделены жирным шрифтом в табл. 1.7. Затем производим расчеты, которые позволяют получить критерий хи-квадрат (табл. 1.7).

Сравниваем $\chi^2_{выч}$ с $\chi^2_{табл}$ при степени свободы $\nu = к - 3 = 6 - 3 = 3$ для $P_{0,95}$. Поскольку $\chi^2_{выч} = 5{,}43 < \chi^2_{табл} = 7{,}815$, теоретическое распределение

частот несущественно отличается от эмпирического, а гипотеза признается состоятельной.

Определим достоверность хи-квадрат по формуле (1.33):

$$D = (\chi^2 - \nu) / \sqrt{2\upsilon} > 3 = (5{,}43 - 3) / \sqrt{2 \cdot 3} = 0{,}99.$$

Таблица 1.7

Сравнение эмпирических и теоретических частот с использованием оценочного критерия Пирсона

Число обследованных механизаторов (классы)	Число фактически больных, φ	Число теоретически больных, φ′	φ – φ′	$(\varphi - \varphi')^2$	$(\varphi - \varphi')^2 / \varphi'$
1–71	1	2			
72–142	3	4 15	–4	16	1,06
143–213	7 11	9			
214–284	10	13	–3	9	0,69
285–355	15	14	1	1	0,07
356–426	12	10	2	4	0,40
427–497	10	11	–1	1	0,09
498–568	8 13	6 8	5	25	3,12
569–639	5	2			
I = 9	$N_1 = 71$	$N_2 = 71$			$\chi^2_{\text{выч}} = \sum 5{,}43$

Полученная величина $D = 0{,}99 < 3$, следовательно нулевая гипотеза признается состоятельной, т. е. влияние природных условий на распространение бронхолегочных заболеваний достоверно.

Глава 2. ДИСПЕРСИОННЫЙ АНАЛИЗ

При планировании эксперимента бывают ситуации, когда исследуемую систему необходимо разбить на группы, отличающиеся между собой в количественном отношении, и установить сходство или различие между ними по влиянию различных факторных величин на признак. Например, определить степень влияния географических условий на ход тех или иных процессов, явлений. Таким условиям лучше всего отвечает дисперсионный анализ, который нашел применение в физической географии.

Дисперсионный анализ позволяет утверждать с определенной долей уверенности наличие влияния на изучаемый объект каждого из условий в отдельности или в их сочетаниях. *Обязательным условием применения дисперсионного анализа является разбивка каждого учитываемого фактора не менее чем на две группы.* Они могут быть представлены как качественными, так и количественными показателями. Качественные показатели приводятся в виде баллов. Анализу подвергаются лишь определяющие поведение объекта факторы, которые установлены исследователем. По количеству определяющих факторов дается название виду дисперсионного анализа (одно-, двух-, трехфакторный и т. д.).

Обработка данных дисперсионного анализа – весьма трудоемкий процесс; облегчает вычисления правильная организация опыта. Порядок расчета в различных видах дисперсионного анализа будет различным, но логическая схема остается единой. Факторы в дисперсионном анализе должны быть независимыми друг от друга; каждый фактор следует разделить на группы, количество которых зависит от поставленной задачи.

Дисперсионный анализ применяется в случаях нормального или близкого к нему распределения выборочных совокупностей. Выборки должны иметь близкие по значению показатели дисперсии σ^2. Количество повторностей в каждой выделенной группе принимается одинаковым.

Основная трудность при использовании дисперсионного анализа – составление комбинационной таблицы для обработки данных *(дисперсионный комплекс).* Если число наблюдений над результативным признаком по отдельным группам изучаемого фактора одинаково, то дисперсионный комплекс называется *равномерным,* если разное, то *неравномерным.* Общее число наблюдений над результативным признаком принято называть *объемом дисперсионного комплекса.*

Порядок действия по каждому виду дисперсионного анализа определяется его основной задачей, которая состоит в делении суммарного или общего варьирования изучаемого признака на доли: варьирование, вызы-

ваемое действием отдельных факторов; варьирование, вызываемое взаимодействием факторов между собой; остаточное варьирование объекта, которое определяется не учитываемыми факторами.

2.1. ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Среди различных видов дисперсионного анализа наиболее часто используется однофакторный. Для выполнения однофакторного анализа в опыте должно быть предусмотрено две повторности и более. Исследуемый фактор разбивается на группы с целью выявления его оптимальной величины, влияющей на результативный признак. Для облегчения расчета можно уменьшить все показатели в пределах дисперсионного комплекса на определенную величину, а затем увеличить конечные результаты на ту же величину.

Географы исследуют не только природные, но и сельскохозяйственные ландшафты (агроландшафты), претерпевающие существенные изменения под воздействием агротехногенеза. Использование системного анализа позволяет не только констатировать изменения в агроландшафте, но и активно включаться в его преобразование.

Известно, что оптимальным условиям питания растений соответствует дерновая легкосуглинистая гумусированная нейтральная почва. Ее можно создать путем внесения в пахотный горизонт добавок минерального грунта определенного механического состава и торфа. Формирование искусственной антропогенной почвы требует полевых экспериментов. В связи с этим поставлена следующая задача: определить влияние на урожай зерна ячменя разных доз торфа (200, 300, 400 т абсолютно сухого вещества на гектар) при внесении его на фоне минеральных, органических удобрений и доломитовой муки. Исходная почва – дерново-подзолистая глееватая связно супесчаная осушенная. После получения сведений об урожайности ячменя в названных условиях составляется таблица дисперсионного комплекса (табл. 2.1), куда заносится исходная информация по группам влияющего фактора (вариантам опыта) и некоторые результаты расчетов (для удобства сделано округление по урожайности до целых чисел). Вначале производим расчет данных по вариантам опыта (строкам).

Результаты разносим по столбцам. Суммарный урожай ячменя по повторностям Σx_i и по каждому варианту опыта вносим в столбец 6 в числителе. Аналогично поступаем с квадратами этих показателей Σx_i^2. Затем в столбце 7 приводим квадраты суммарного урожая ячменя по повторностям $(\Sigma x_i)^2$. Затем вычисляем среднее арифметическое M_i по каждому варианту опыта, заносим в столбец 8, вычисляем $M_{\text{общ}}$.

Таблица 2.1

Однофакторный дисперсионный анализ

<table>
<tr><td colspan="2" rowspan="2">Варианты опыта (фактор)</td><td colspan="4" rowspan="2">Урожай ячменя по повторностям, ц/га*</td><td colspan="3">По повторностям (признакам) (i)</td></tr>
<tr><td>$\frac{\sum x_i}{\sum (x_i^2)}$</td><td>$\left(\sum x_i\right)^2$</td><td>$M_i$</td></tr>
<tr><td colspan="2">Контроль (фон)</td><td>$\frac{20}{400}$</td><td>$\frac{21}{441}$</td><td>$\frac{22}{484}$</td><td>$\frac{20}{400}$</td><td>$\frac{83}{1725}$</td><td>6889</td><td>20,75</td></tr>
<tr><td colspan="2">Фон +200 т/га торфа</td><td>$\frac{30}{900}$</td><td>$\frac{32}{1024}$</td><td>$\frac{32}{1024}$</td><td>$\frac{31}{961}$</td><td>$\frac{125}{3909}$</td><td>15 625</td><td>31,25</td></tr>
<tr><td colspan="2">Фон + 300 т/га торфа</td><td>$\frac{35}{1225}$</td><td>$\frac{36}{1296}$</td><td>$\frac{35}{1225}$</td><td>$\frac{36}{1296}$</td><td>$\frac{142}{5042}$</td><td>20 164</td><td>35,50</td></tr>
<tr><td colspan="2">Фон + 400 т/га торфа</td><td>$\frac{36}{1296}$</td><td>$\frac{35}{1225}$</td><td>$\frac{37}{1369}$</td><td>$\frac{37}{1369}$</td><td>$\frac{145}{5259}$</td><td>21 025</td><td>36,25</td></tr>
<tr><td rowspan="4">По факторам (k)</td><td>$\sum x_k$</td><td>121</td><td>124</td><td>126</td><td>124</td><td rowspan="2">$\sum\sum x_{i,k}$ 495
$\sum\left(\sum x_{i,k}^2\right)$ 15935</td><td rowspan="2">$\sum\left(\sum x_i\right)^2$ 63703</td><td rowspan="3"></td></tr>
<tr><td>$\sum (x_k^2)$</td><td>3821</td><td>3986</td><td>4102</td><td>4026</td></tr>
<tr><td>$\left(\sum x_k\right)^2$</td><td>14 641</td><td>15 376</td><td>15 876</td><td>15 376</td><td></td></tr>
<tr><td>M_k</td><td>30,25</td><td>31,00</td><td>31,50</td><td>31,00</td><td colspan="2">$\sum\left(\sum x_k\right)^2 = 61\ 269$</td><td>$M_{\text{общ}}$ 30,93</td></tr>
</table>

Примечание. * В числителе – опытные данные, в знаменателе – квадраты этих показателей.

После получения данных по вариантам опыта производим расчет необходимых показателей по повторностям (x_k). Сначала суммируем данные урожайности ячменя и приводим в строке под чертой Σx_k. Суммы сумм урожайности ячменя по вариантам опыта и повторностям должны совпасть и дать сумму всех вариант ($\Sigma\Sigma x_{i,k} = 495$). Аналогично суммируем квадраты этих показателей по повторностям (Σx_k^2). Суммы сумм квадратов по вариантам и повторностям опыта должны совпасть и дать сумму квадратов всех вариант ($\Sigma x_i^2 = \Sigma x_k^2 = 15\ 935$). Ниже вписываем результаты возведения в квадрат сумм вариант по каждой повторности $(\Sigma x_k)^2$ и суммируем их: $\Sigma(\Sigma x_k)^2 = 61\ 269$. Вычисляем средние арифметические по каждой повторности опыта M_k. Общее среднее арифметическое всех вариант опыта составляет $M_{\text{общ}} = (\Sigma x_{i,k}) / N = 495 : 16 = 30{,}93$.

Следующий этап работы – нахождение сумм квадратов отклонений, т. е. расчленение общего варьирования признака на составные части исходя из равенства:

$$\Theta = \Theta_1 + \Theta_2 + \Theta_3,$$

где Θ – сумма квадратов отклонений по общему варьированию данных, Θ_1 – по группам фактора (варианты опыта), Θ_2 – по повторностям опыта, Θ_3 – по остаточному варьированию, вызванному неучтенными факторами.

Общая сумма квадратов отклонений вычисляется:

$$\Theta = \Sigma(\Sigma x^2_{i,k}) - (\Sigma\Sigma x_{i,k})^2 / N.$$

Подставив данные из табл. 2.1, получим: Θ =15 935 – 495^2 : 16 = 621. Затем находим сумму квадратов отклонений по группам фактора (варианты опыта) по формуле:

$$\Theta_1 = [\Sigma(\Sigma x_i)^2 - (\Sigma\Sigma x_{i,k})^2/k] / i, \qquad (2.1)$$

где k – число групп фактора, т. е. 4; i – число повторностей, т. е. 4. В данном случае должно выдержаться равенство $N = ki = 4 \cdot 4 = 16$. По формуле (2.1) вычислим:

$$\Theta_1 = [63703 - 495^2 : 4] : 4 = 611{,}75.$$

Сумму квадратов отклонений по повторностям опыта находим по формуле

$$\Theta_2 = [\Sigma(\Sigma x_k)^2 - (\Sigma\Sigma x_{i,k})^2/i] / k, \qquad (2.2)$$

где i – число повторностей, т. е. 4; k – число слагаемых в каждой сумме Σx_k, т. е. 4.

Вычисляем Θ_2 по формуле (2.2):

$$\Theta_2 = [61269 - 495^2 : 4] : 4 = 3{,}25.$$

Сумма квадратов отклонений по остаточному варьированию определяется из равенства

$$\Theta_3 = \Theta - \Theta_1 - \Theta_2. \qquad (2.3)$$

Подставив значение вычисленных сумм соответствующих квадратов отклонений в формулу (2.3), получим

$$\Theta_3 = 621 - 611{,}75 - 3{,}25 = 6{,}00.$$

Проводим дисперсионный анализ данных урожая ячменя (табл. 2.2). Вносим в таблицу рассчитанные суммы квадратов отклонений (Θ, Θ_1, Θ_2, Θ_3). Число степеней свободы получаем следующим образом: по общей сумме квадратов отклонений $\nu = N - 1 = 16 - 1 = 15$; по вариантам опыта $\nu_1 = n_1 - 1 = 4 - 1 = 3$; по повторностям $\nu_2 = n_2 - 1 = 4 - 1 = 3$; по остаточной сумме $\nu_3 = \nu - \nu_1 - \nu_2 = 15 - 3 - 3 = 9$.

Дисперсия определяется путем деления сумм квадратов отклонений (Θ, Θ_1, Θ_2, Θ_3) на соответствующие им числа степеней свободы (ν, ν_1, ν_2, ν_3), что можно выразить в общем виде формулой $\sigma^2 = \Theta/\nu$, получим $\sigma^2 = 621 : 15 = 41{,}40$.

Таблица 2.2

Результаты однофакторного дисперсионного анализа

Варьирование данных	Сумма квадратов отклонений, Θ	Степень свободы, ν	Дисперсия, $\sigma^2 = \Theta / \nu$	Критерий Фишера	
				$F_ф$	$F_т$
Общее по опыту	621,00	15	41,40	–	–
По вариантам опыта	611,75	3	203,91	304,31	8,81
По повторностям	3,25	3	1,08	1,61	8,81
Случайное (остаточное)	6,00	9	0,67	–	–

Оценку сходства или различия между вариантами опыта можно проводить по критерию Фишера, критерию Стьюдента или НСР.

Поскольку $F_ф > F_т$ (см. табл. 2.2 и прил. 5), то это позволяет сделать вывод, что внесение больших доз торфа положительно влияет на величину урожая ячменя в агроландшафте.

Наиболее распространен в дисперсионном анализе для оценки результатов опыта критерий НСР, алгоритм которого приводим ниже. Вначале определяем среднее квадратическое отклонение из дисперсии, полученной в результате случайного варьирования (см. табл. 2.2): $\sigma = \sqrt{\sigma_3^2}$, затем вычисляем обобщенную ошибку среднего: $m_M = \sigma / \sqrt{N_{повт}}$. Поскольку ошибка среднего для всех сравниваемых вариант одна и та же, формула для расчета ошибки разности может быть преобразована: $m_d = \sqrt{2m^2}$. Наименьшую существенную разность рассчитываем по формуле (1.24). Используя исходные данные, вычислим НСР по указанному выше алгоритму:

$$\sigma = \sqrt{0,67} = 0,82;\ m_M = 0,82 / \sqrt{4} = 0,41;$$

$$m_d = \sqrt{2 \cdot 0,41^2} = 0,58;\ \text{НСР}_{0,95} = 0,58 \cdot 2,26 = 1,31;$$

$$\text{НСР}_{0,99} = 0,58 \cdot 3,25 = 1,88.$$

Из полученных результатов дисперсионного анализа вытекает следующий вывод (табл. 2.3). Величина $\text{НСР}_{0,95}$ и $\text{НСР}_{0,99}$ меньше величины прибавки урожая зерна ячменя, поэтому внесение высоких доз торфа положительно влияет на урожай. Лучший результат получен в варианте с дозой внесения торфа 400 т/га, где прибавка зерна ячменя составила 15,5 ц/га.

В случае необходимости можно рассчитать ошибки частных средних арифметических по повторностям:

$$m_п = \sqrt{\sigma_{ост}^2 / N_n} = \sqrt{0,67 / 4} = 0,41;$$

Таблица 2.3

Влияние высоких доз торфа на урожай ячменя

Вариант опыта	Урожай ячменя по повторностям				Среднее	Прибавка
Контроль (фон)	20	21	22	20	20,75	–
Фон + 200 т/га	30	32	32	31	31,25	10,50
Фон + 300 т/га	35	36	35	36	35,50	14,75
Фон + 400 т/га	36	35	37	37	36,25	15,50
$НСР_{0,95}$, ц/га $НСР_{0,99}$, ц/га p	1,31 1,88 1,32 %					

по вариантам опыта:

$$m_{в} = \sqrt{\sigma^2_{ост} / N_{в}} = \sqrt{0,67 / 4} = 0,41.$$

Ошибку общего среднего арифметического используют для вычисления точности опыта. Показатель точности опыта для общего среднего арифметического вычисляется следующим образом:

$$p_{Mобщ} = (m_{общ} / M_{общ}) \cdot 100 = (0,41 : 30,90) \cdot 100 = 1,32\ \%.$$

Поскольку p = 1,32 %, т. е. < 3 %, то опыт признается достаточно точным.

Аналогичным образом вычисляется точность опыта для частных средних арифметических по вариантам опыта и по повторностям:

$$p_{в} = (m_{в} / M_{в}) \cdot 100; \quad p_{п} = (m_{п} / M_{п}) \cdot 100.$$

Глава 3. КЛАСТЕРНЫЙ АНАЛИЗ

При проведении землеустроительных исследований, как правило, возникает проблема *объединения по сходству* (*кластеризация*) объектов, которые характеризуются множеством признаков, выраженных в разных единицах измерения. Для этой цели используется *кластерный анализ*. Поскольку кластерный анализ занимается классификацией объектов, а факторный исследует связи между ними, то оба метода дополняют друг друга и между ними иногда трудно провести четкие границы.

Методологические особенности кластерного анализа сводятся к выявлению единой меры, охватывающей ряд исследуемых признаков. Эти признаки объединяются с помощью метрики (расстояния) в один кластер сходства группируемых объектов.

Состояние любого объекта может быть описано с использованием *многомерного признака*, или *многомерной случайной величины* (x_1, x_2,…, x_n). Примером количественных признаков при зонировании территории города может служить площадь строений (x_1), количество исторических памятников (x_2), количество промышленных предприятий (x_3) и т. д. Их можно объединить в один качественный признак – инфраструктурные условия города. Таким образом, состояние любого объекта может быть описано с помощью многомерного признака.

Исследование нескольких аналогичных объектов (городов) обязывает проводить разбиение совокупности объектов на однородные группы, т. е. провести классификацию городов по сходству признаков (x_1, x_2 …). В зависимости от специальности и природы используемых методов исследователи называют классификацию многомерных наблюдений как *распознавание образов с учителем* (численной таксономией), *кластер-анализом без учителя, дискриминантным* анализом.

Таксономические методы классификации объектов основываются на выделении групп объектов наиболее близких в многомерном пространстве. Для определения степени сходства объектов вычисляются таксономические расстояния между ними. Если исследователь имеет перед собой образы будущих групп – обучающие выборки, то группировка выполняется методом дискриминантного анализа. При отсутствии обучающих выборок используется кластерный анализ (В. В. Глинский, В. Г. Ионин, 1998). В отличие от дискриминантного анализа (С. А. Айвазян и др., 1984), отсутствие классифицированных обучающих выборок в кластерном анализе значительно усложняет решение задачи классификации.

Для оценки сходства объектов по ряду признаков используют три типа мер:

● *коэффициент подобия* – для группировки объектов и признаков, если уровни показателей являются действительно целыми числами;

● *коэффициенты связи* – чаще применяются для группировки признаков с использованием коэффициента корреляции;

● *показатели расстояния* – характеризуют степень взаимной удаленности признаков и применяются в основном для кластеризации объектов; признаки объектов должны быть независимыми, что предварительно можно уточнить с помощью корреляционного анализа.

Многомерное наблюдение может быть интерпретировано геометрически в виде точки в многомерном пространстве. Геометрическая близость точек в пространстве означает близость физических состояний объектов, их однородность. Решающим в интерпретации остается выбор масштаба метрики, т. е. задание расстояния между объектами, которые объединяют или разъединяют объекты. В результате разбиения объектов на группы по сходству признаков образуются *кластеры* (*таксоны, образы*).

Выбор метрики (меры близости) является важнейшим моментом исследования, от которого зависит окончательный вариант разбиения объектов на группы. Это зависит от цели исследования, физической и статистической природы вектора наблюдений (x), полноты априорных сведений о характере вероятностного распределения x.

В задачах кластер-анализа широко используются следующие метрики: Эвклида, Махаланобиса, Хемминга, меры близости, задаваемые потенциальной функцией. Эвклидова метрика наиболее употребительна.

Обычно среднее эвклидово расстояние рассчитывается по формулам:

$$d_{kl} = \left[\frac{1}{m} \sum_{j=1}^{m} \left(Z_{k_j} - Z_{l_j} \right)^2 \right]^{1/2}, \qquad (3.1)$$

где m – число признаков x; Z_{k_j}, Z_{l_j} – стандартизированные значения признака j для k и l объектов соответственно, или:

$$d_{kl} = \sqrt{\frac{\left(Z_{k_{j1}} - Z_{l_{j1}} \right)^2 + \left(Z_{k_{j2}} - Z_{l_{j2}} \right)^2 + \ldots + \left(Z_{k_{jn}} - Z_{l_{jn}} \right)^2}{m}}.$$

Если не учитывать число признаков x – m, формула примет вид:

$$d_{kl} = \left[\sum_{j=1}^{m} (Z_{kj} - Z_{lj})^2 \right]^{1/2}. \qquad (3.2)$$

Последняя формула (3.2) менее объективна, так как не учитывает число признаков, количество которых может изменяться от трех и более.

Расчет упрощается, если в качестве метрики использовать l_1-*норму*:

$$d_{kl} = \sum_{j=1}^{m} (Z_{kj} - Z_{lj}). \qquad (3.3)$$

Эти метрики используются в следующих случаях:

● наблюдения x извлекаются из генеральных совокупностей, описываемых многомерным нормальным законом с ковариационной матрицей (совместное изменение двух признаков), где компоненты x взаимно независимы и имеют одинаковую дисперсию;

● компоненты $x_1, x_2, \ldots, x_p$ вектора наблюдений x однородны по своему физическому смыслу и все важны;

● факторное пространство совпадает с геометрическим; понятие близости объектов соответственно совпадает с понятием геометрической близости в этом пространстве.

«Взвешенное» эвклидово расстояние определяется:

$$d_{kl} = \sqrt{\omega_1 \left(Z_{k_{j1}} - Z_{l_{j1}}\right)^2 + \omega_2 \left(Z_{k_{j2}} - Z_{l_{j2}}\right)^2 + \ldots + \omega_n \left(Z_{k_{jn}} - Z_{l_{jn}}\right)^2}, \qquad (3.4)$$

или

$$d_{kl} = \sqrt{\left(Z_k - Z_l\right)' \wedge' \sum^{-1} \wedge \left(Z_k - Z_l\right)},$$

где $\sum$ – ковариационная матрица генеральной совокупности, из которой извлекаются наблюдения Z; $\wedge$ – некоторая симметричная неотрицательно-определенная матрица «весовых» коэффициентов λ_{mq}, которая чаще всего выбирается диагональной; ’ – штрих-символ операции транспонирования вектора; –1 – обращение матрицы $\sum$.

Хемминга расстояние используется как мера различия объектов, задаваемых дихотомическими признаками (деление объекта на две составляющие):

$$d_{kl} = \sum_{j=1}^{m} \left| Z_{kj} - Z_{lj} \right|. \qquad (3.5)$$

Обычно признаки заданы в виде набора нулей и единиц: 0, если $a_i = b_i$; 1, если $a_i \neq b_i$.

Средним линейным отклонением оценивается расстояние между объектами по Хеммингу:

$$d_{kl} = \frac{1}{m}\sum_{j=1}^{m}\left|Z_{k_j} - Z_{l_j}\right|. \qquad (3.6)$$

Таким образом, при решении задач классификации могут быть использованы разные меры сходства между объектами. Выбор метрики зависит от вида информации, характеризующей объекты в пространстве признаков и требует тщательного критического анализа.

Этапы работ в кластерном анализе

Решение задач классификации объектов с использованием кластерного анализа проводится в определенной последовательности. Многомерный анализ делится на три этапа:

- составляется таблица исходной информации с указанием объектов и их признаков;
- проводится нормализация исходной информации с использованием среднего квадратического отклонения;
- по нормализованным данным рассчитывается метрика, строится дендрограмма и проводится содержательная интерпретация полученных результатов.

На первом этапе при формировании таблицы выбор объекта зависит от места и масштаба исследования. Каждый объект должен быть пространственно локализован и одного ранга (уровня). Показатели должны отражать существенные черты или свойства исследуемых объектов и характеризовать их всесторонне.

На втором этапе нормализация значений исходных показателей по объектам проводится потому, что исходные данные выражены обычно в разных единицах измерения и между ними проводить арифметические действия невозможно без перевода их в безразмерные единицы.

Наиболее распространенный способ нормализации показателей проводится с использованием среднего квадратического отклонения по формуле:

$$\hat{Z}_{ij} = (Z_{ij} - \overline{Z_{ij}})/\sigma_j; \qquad (3.7)$$

$$\sigma_j = \sqrt{\frac{\sum\left(Z_{ij} - \overline{Z}_{ij}\right)^2}{N_j}}, \qquad (3.8)$$

где $\hat{Z}_{ij}$ – нормализованная безразмерная величина; Z_{ij} – индивидуальные значения по столбцам матрицы; $\overline{Z}_{ij}$ – среднее значение по столбцам мат-

рицы; σ_j – среднее квадратическое отклонение по столбцам; N_j – объем выборки по столбцам.

Составляется матрица нормализованных показателей.

На третьем этапе по нормализованным показателям рассчитывается метрика по одному из предложенных выше способов, учитывая условия задачи. Классификацию объектов производят приемами таксономического или факторного анализа.

При количестве координат (показателей) в многомерном пространстве более трех графически интерпретировать таксономические расстояния невозможно. Поэтому таксономические расстояния определяют на основе функции расстояний. Чаще всего используется эвклидова метрика.

На основе матрицы таксономических расстояний производится группировка объектов с использованием разных приемов, из них наиболее распространенные – метод дендритов, вроцлавская таксономия, дендродерево Берри.

Метод дендритов. Объекты, разделенные на кластеры, можно изобразить в виде дендрограммы, которая представляет собой графическое изображение матрицы расстояний или сходства. Такой анализ объектов исследования носит название метода дендритов.

Представим дендрограмму с шестью объектами ($n = 6$) (рис. 3.1). Объекты 1 и 3 наиболее близки, т. е. наименее удалены друг от друга, поэтому объединяются в один кластер на уровне сходства, равном 0,9 (образуют 1-й шаг). Объекты 4 и 5 объединяются при уровне сходства 0,8 (2-й шаг). На 3-м и 4-м шагах процесса образуются кластеры 1, 3, 6 и 5, 4, 2, соответствующие уровню сходства соответственно 0,7 и 0,6. Окончательно все объекты группируются в один кластер при уровне сходства 0,5.

Вид дендрограммы зависит от выбора меры сходства или расстояния и метода кластеризации. Например, разработаны алгоритмы кластерного анализа, позволяющие проводить классификацию (группировку) многомерных наблюдений (строк и столбцов матрицы x) с помощью следующих мер сходства: выборочного коэффициента корреляции, модуля выборочного коэффициента корреляции, косинуса угла между векторами, модуля косинуса угла между векторами, эвклидова расстояния и т. д.

Вроцлавская таксономия. По матрице таксономических метрик (табл. 3.1) строится граф-дерево, вершинами которого будут объекты группировки.

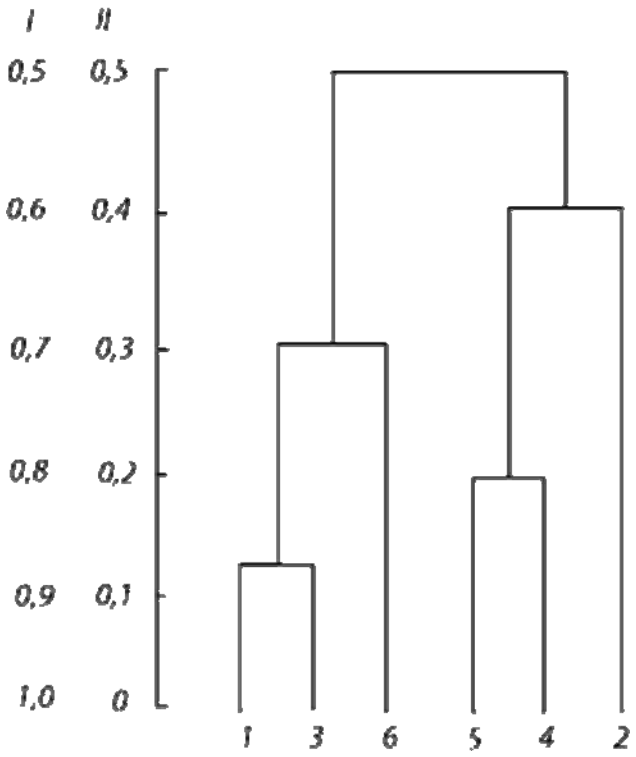


Рис. 3.1. Общий вид дендрограммы:
I – сходство, *II* – расстояние

Таблица 3.1

Матрица таксономических метрик

Объекты	A	B	C	D	E	F	G	H	I	J
A	0	**1,15**	5,05	4,22	3,54	3,30	**2,56**	3,62	3,10	**1,67**
B	**1,15**	0	6,41	4,53	3,81	3,84	2,99	4,53	3,88	2,63
C	5,05	6,41	0	4,04	4,82	4,06	4,83	**3,07**	4,34	4,14
D	4,22	4,53	4,04	0	**1,66**	1,68	2,34	2,80	2,99	4,02
E	3,54	3,81	4,82	**1,66**	0	**0,96**	1,34	2,76	2,26	3,72
F	3,30	3,84	4,06	1,68	**0,96**	0	**1,11**	1,80	1,51	3,22
G	**2,56**	2,99	4,83	2,34	1,34	**1,11**	0	2,24	**1,38**	3,01
H	3,63	4,53	**3,07**	2,80	2,76	1,80	2,24	0	**1,33**	3,09
I	3,10	3,88	4,34	2,99	2,26	1,54	**1,38**	**1,33**	0	3,18
J	**1,67**	2,63	4,14	4,02	3,76	3,22	3,01	3,09	3,18	0

Порядок построения графа следующий (рис. 3.2). В каждом столбце или ряде зеркальной матрицы (по диагонали нули) находится минимальная величина метрики. Вначале откладывается в выбранном масштабе наименьшая среди метрик в матрице между объектами (EF = 0,96). Затем последовательно к отложенным объектам откладываем минимальные метрики других столбцов-объектов: FG = 1,11, ED = 1,66, GI = 1,38, IH = 1,36, HC = 3,07,

GA = 2,56, AB = 1,15, AJ = 1,67. Метрика используется только один раз. Если при построении графа на нем образуется замкнутый цикл, то замыкающее ребро цикла во внимание не принимается и вместо него откладывается ребро, которое отвечает другой минимальной метрике в данном столбце матрицы.

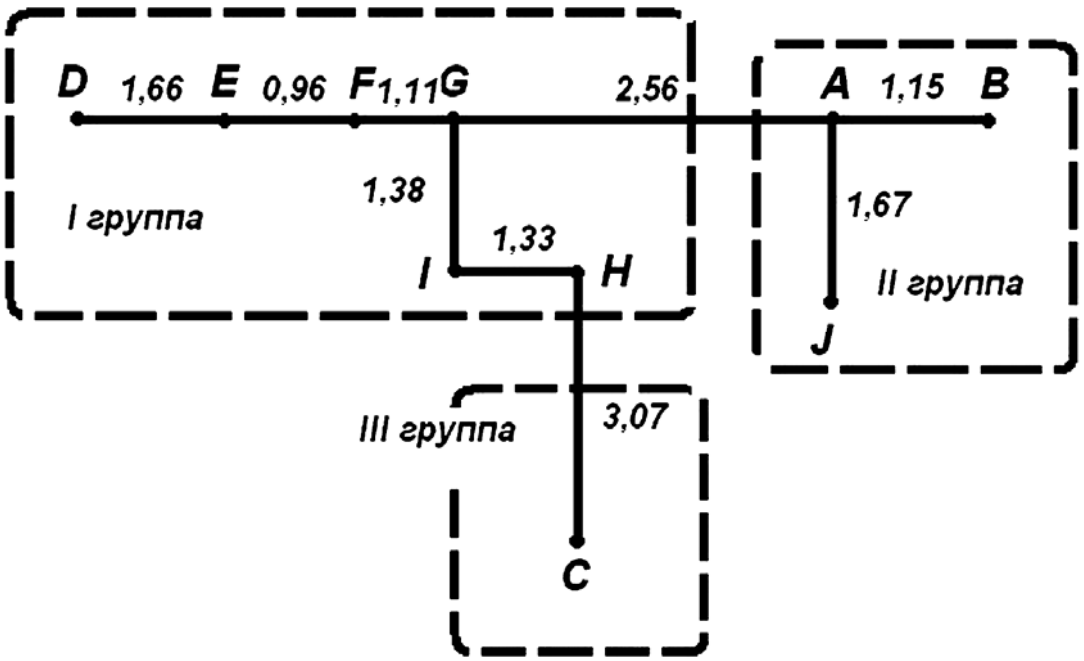


Рис. 3.2. Вроцлавский дендрит

После построения графа с нанесением всех объектов проводится группировка (классификация) объектов. Задается определенная величина таксономической метрики, которая является основой классификации. Таким образом, граф разбивается на подграфы, в пределах которых объекты должны располагаться компактно (близко друг к другу) (рис. 3.2). В конце дается интерпретация полученных результатов с учетом исходной таблицы первоначальных данных. Чем меньшая величина метрики объединяет объекты на графе, тем более близкие по своим значениям исходные показатели в этих объектах.

Метод дендро-дерева Б. Берри. В матрице таксономических метрик выбирается наименьший элемент, который связывает два объекта (см. табл. 3.1): EF = 0,96. Метрика свидетельствует, что объекты E и F находятся на минимальном и одинаковом расстоянии по отношению к другим объектам. Поэтому их можно заменить одним, присвоив символ M (рис. 3.2).

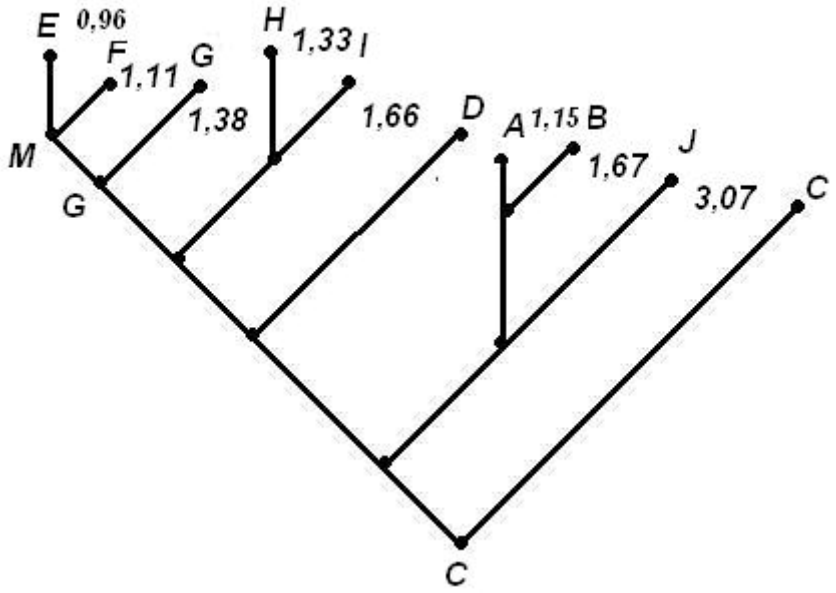


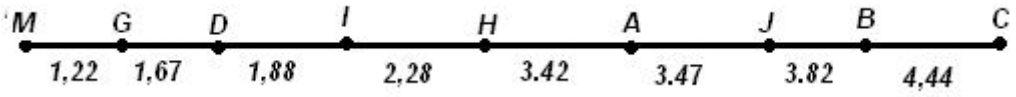
Рис. 3.3. Дендро-дерево Берри

В дальнейшем на горизонтальной линии размещаем объекты последовательно по мере увеличения их метрик с учетом связи с первыми объектами EF. Объект G связывается с F метрикой 1,11, объект I с G – 1,38, I с H – 1,33, E с D – 1,66. Далее связь неотложенных объектов (A, B, J, C) с отложенными прерывается. В таких случаях внутри этих объектов ищем наименьшие метрики между ними: A и B связывает минимальная метрика 1,15; A и J – 1,67. Объект C связан наименьшей метрикой 3,07 с ранее отложенной H, поэтому он выделяется самостоятельно в конце по прямой линии (рис. 3.2).

Отложенные объекты на горизонтальной линии с минимальными метриками связываются между собой (H и I: A и B) или выделяется самостоятельно с общей наклонной линией M-C, на которой откладываются вычисленные метрики от объекта M (E-F) путем вычисления усредненных величин, используя данные матрицы по строкам E-F (табл. 3.1):

A = (3,54 + 3,30) / 2 = 3,42; B = (3,81 + 3,84) / 2 4 = 3,82;
C = (4,82 + 4,06) / 2 = 4,44; D = (1,66 + 1,68) / 2 = 1,67;
G = (1,34 + 1,11) / 2 = 1,22; H = (2,76 + 1,80) / 2 = 2,28;
I = (2,26 + 1,51) / 2 = 1,88; J = (3,72 + 3,22) / 2 = 3,47.

Располагаем объекты относительно M по возрастающей величине на линии и производим группировку:



Расчленение графа на подгруппы для определения количества групп объектов может производится в процессе его построения (рис. 3.3): EF; HI; ABJ.

При делении объектов на классы важным критерием является минимизация внутригрупповой и максимизация межгрупповой дисперсии. Практически количество классов определяется априорно, т. е. по внешнему виду дендро-дерева. В выделенном классе объекты по анализируемым признакам являются сходными (однородными). Если они соседние в пространстве, то образуют однородный регион.

В нашем примере объекты можно объединить в 4 класса (EFG; HID; ABJ; C) по минимальным метрикам между объектами и по усредненным относительно объекта M (E, F).

Пример кластерного анализа по способу Вроцлавский дендрит

Задача: провести зонирование территории города по предложенным признакам.

Этапы работы:

1. Подсчитываем сумму, среднее и сигму по столбцам:

	$\sum$	среднее	σ
1-й столбец	10,6	1,8	1,2
2-й столбец	115	19,2	6,6
3-й столбец	21	3,5	2,14
4-й столбец	48	8	5,1
5-й столбец	22	3,7	1,97
6-й столбец	15,2	2,5	1,64
7-й столбец	255	42,5	19,1

2. Трансформируем количественные показатели в числа без измерений (табл. 3.3) с использованием формулы (3.7, 3.8).

Таблица 3.3

Количественные показатели для зонирования города

город Минск	Площадь застройки, га		Количество исторических памятников	Количество архитектурных памятников	Количество промышленных предприятий	Площадь лесной зоны, га	Шумовое загрязнение, балл
	дерев.	бетон.					
объект № 1	0,1	25	5	10	2	2	80
объект № 2	0,5	10	7	12	3	3	40
объект № 3	1,5	15	3	16	5	0,5	30
объект № 4	2,0	17	4	5	4	0,7	50
объект № 5	3,0	18	1	4	7	5	20
объект № 6	3,5	30	1	1	1	4	35

3. Рассчитываем расстояния (метрику) между объектами по формуле (3.3) и проставляем в матрицу ниже:

	1	2	3	4	5	6
1	0	3,36	3,99	3,11	5,62	4,10
2	3,36	0	3,01	3,01	4,52	5,40
3	3,99	3,01	0	2,32	5,21	5,67
4	3,11	3,01	2,32	0	3,83	3,90
5	5,62	4,52	5,21	3,83	0	3,76
6	4,10	5,40	5,67	3,90	3,76	0

4. По полученным расстояниям (метрикам) по столбцам или строкам выбираем наименьшие расстояния и откладываем их в масштабе до тех пор, пока не нанесем все объекты (рис. 3.4).

Таблица 3.4

Нормализованные безразмерные данные

1	–1,42	0,88	0,7	0,39	–0,86	–0,31	1,96
2	–1,08	–1,4	1,63	0,78	–0,36	0,31	–0,13
3	–0,25	–0,64	–0,23	1,56	0,66	–1,25	–0,65
4	0,17	–0,33	0,23	–0,58	0,15	–1,12	0,4
5	1	–0,18	–1,16	–0,78	1,68	1,56	–1,18
6	1,42	1,64	–1,16	–1,37	–1,37	0,93	–0,39

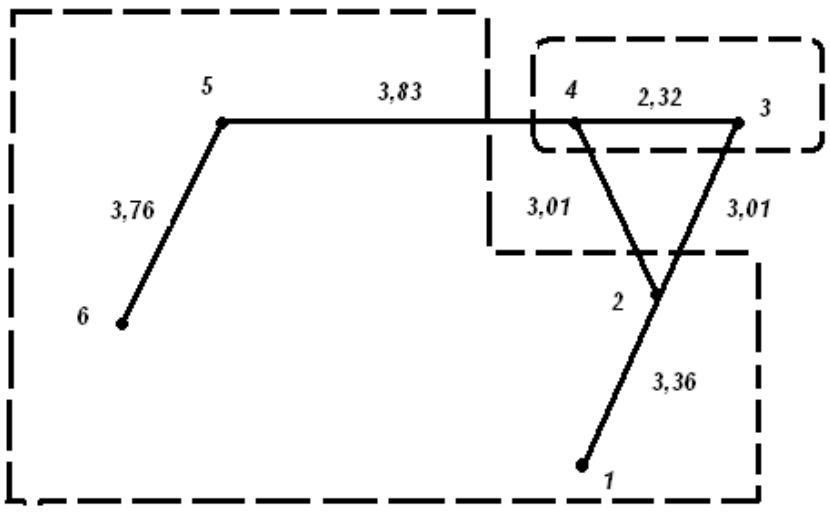


Рис. 3.4. Кластеризация объектов города
(Вроцлавский дендрит)

5. Первое расстояние выбирается наименьшим (2,32) в матрице между объектами 3 и 4, затем метрику 3,1 между объектами 3 и 2 и т. д. (рис. 3.4). Продолжаем работу по построению графика до тех пор, пока не отложим все объекты.

6. Выбираем масштаб для разделения метрик (0–1, 1,1–2, 2,1–3, 3,1–4) с целью проведения классификации и объединяем объекты в одну группу по выбранному масштабу. Эти объекты будут иметь близкие значения по большинству показателей, которыми характеризуются объекты.

В приведенном примере выделено две группы. В первую группу вошли объекты 3 и 4 с минимальным расстоянием (метрикой 2,3) между ними; вторую группу образовали объекты 1, 2, 5, 6 с метрикой разделения более 3.

Глава 4. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Термин «корреляция» означает соотношение, соответствие. Представление о корреляции как о взаимозависимости случайных переменных величин лежит в основе статистической теории корреляции – изучение зависимости вариации признака от окружающих условий. Одни признаки выступают в роли *влияющих (факторных)*, другие – на которые влияют, *результативных*. Зависимости между признаками могут быть *функциональными и корреляционными.* Функциональные связи характеризуются полным соответствием между изменением факторного признака и изменением результативной величины. Каждому значению признака-фактора соответствует определенное значение результативного признака. В корреляционных связях между изменением факторного и результативного признака нет полного соответствия. В сложном взаимодействии находится сам результативный признак. Поэтому результаты корреляционного анализа имеют значение в данной связи, а интерпретация этих результатов в общем виде требует построения системы корреляционных связей. Они характеризуются множеством причин и следствий, с их помощью устанавливается тенденция изменения результативного признака при изменении величины факторного признака. Например, на производительность труда влияют факторы степени совершенствования техники и технологии, уровень механизации и автоматизации труда, специализации производства, состав работающих, текучесть кадров и т. д.

В природе и обществе явления и события протекают по характеру корреляционной связи, когда при изменении величины одного признака существует тенденция изменения другого признака. Корреляционная связь – это частный случай статистической связи. *Корреляционный анализ используется при установлении тесноты зависимости между явлениями, процессами, объектами.*

С помощью корреляции можно дать лишь формальную оценку взаимосвязей. Поэтому прежде чем приступать к вычислению коэффициентов корреляции между любыми признаками, следует теоретически установить, имеется ли между этими признаками взаимосвязь. Ведь формально статистика может доказать несуществующие связи, например, между высотой здания в городе и урожайностью пшеницы в фермерских хозяйствах.

Связь между явлениями (корреляция) определяется путем постановки опытов, статистического анализа. Корреляцию не следует отождествлять с причинностью. Однако необходимо иметь в виду, что доказательство математической связи должно опираться на реальную зависимость между явлениями. Любой показатель связи служит приближенной оценкой рассматриваемой зависимости и не является гарантией существования жесткой (функ-

циональной) соподчиненности. Отсутствие жесткой зависимости в природе и обществе способствует саморегуляции процессов, явлений, систем.

По направлению связь может быть *прямой* и *обратной*; по характеру – *функциональной* или *статистической (корреляционной)*; по величине – *слабой, средней* или *сильной*; по форме – *линейной* и *нелинейной*; по количеству коррелируемых признаков – *парной и множественной*.

Функциональная зависимость характерна для геометрических форм, технических систем, когда каждому значению одного признака соответствует точное значение другого. Это пример взаимосвязи площади прямоугольника и длины его одной из сторон. Такая зависимость полная или исчерпывающая.

Если на признак влияет несколько факторов, то приходится оценивать множественную корреляцию. *Множественная корреляция* служит основой выявления связей между признаками, но требует строгой нормальности и прямолинейности распределения, поэтому использование ее может быть затруднено. Изучение влияния главных факторов на признак более детально и точно исследуют путем факторного анализа.

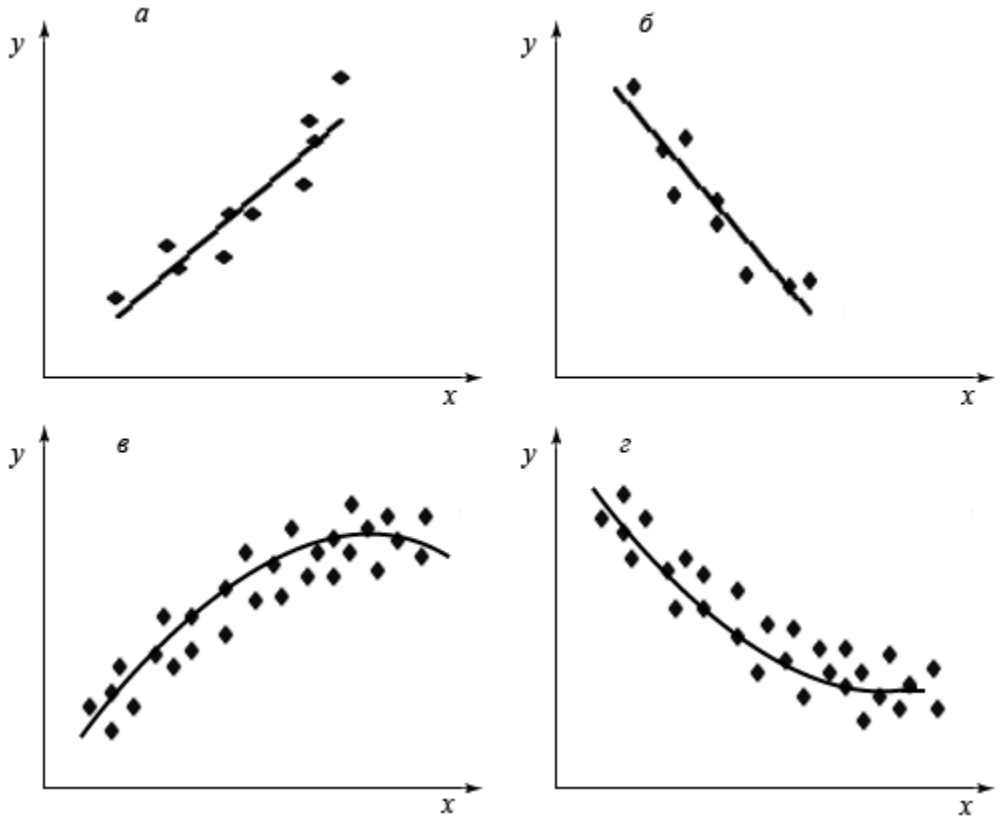


Рис. 4.1. Форма корреляционной связи:
а – прямая линейная; *б* – обратная линейная;
в – параболическая; *г* – гиперболическая

В практической работе по установлению корреляции между признаками и явлениями необходимо придерживаться следующей последовательности:

– на основании проведенных исследований предварительно определяют, существует ли связь между рассматриваемыми признаками;

– если связь между ними существует, устанавливают ее форму, направление и тесноту по построенному графику.

Корреляционный анализ решает следующие задачи:

- Установление направления и формы связи.
- Оценка тесноты связи.
- Оценка репрезентативности статистических оценок взаимосвязи.
- Установление величины детерминации (доли взаимовлияния) коррелируемых факторов.

Для оценки связи используют коэффициент корреляции (r) при линейной зависимости, корреляционное отношение (η) при нелинейной зависимости.

4.1. ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ

Для установления формы зависимости по исходным вариантам (x, y) строится график. В случае линейной зависимости вычисляется коэффициент корреляции (r), при нелинейной – корреляционное отношение (η). В зависимости от величины разброса точек на графике можно предварительно установить форму (см. рис. 4.1) и тесноту (см. рис. 4.2) связи.

Линия регрессии по координатам точек на графике проводится таким образом, чтобы точки в равном количестве находились по обе стороны линии.

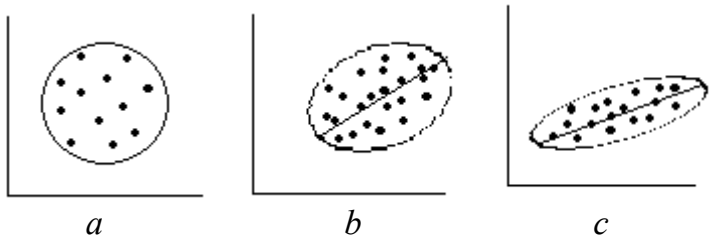


Рис. 4.2. Степень рассеяния частот и величина связи:
$a - r \approx 0$; $b - r \approx 0{,}5$; $c - r \approx 0{,}7$

Точное значение r получаем расчетным способом как при прямой (r от 0 до 1), так и при обратной (r от 0 до –1) зависимости следующим образом:

$$r = \frac{\sum (x_i - M_x)(y_i - M_y)}{\sqrt{\sum (x_i - M_x)^2 \sum (y_i - M_y)^2}}, \qquad (4.1)$$

где $(x_i - M_x)$, $(y_i - M_y)$ – отклонения значений индивидуальных вариант x_i и y_i от их средних значений M_x и M_y.

Более простой алгебраический расчет коэффициента вариации с учетом объема выборки (n):

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right)}}. \qquad (4.2)$$

Исходные данные и суммы по ним получаем из представленной формы:

x_i	x_i^2	$x_i - M_x$	y_i	y_i^2	$y_i - M_y$	xy	$(x_i - M_x)(y_i - M_y)$

Принимается следующая характеристика тесноты корреляционной связи: если r (η) = 0 ± 0,4, то связь считается слабой; от ±0,4 до ±0,7 – средняя; от ±0,7 до ±1 – сильная; r = ±1 и η = 1 – связь считается функциональной.

Достоверность вычисленного коэффициента корреляции может быть установлена двумя путями: путем сравнения с табличным значением r (прил. 7); второй путь – через критерий Стьюдента. Если $r_{выч} > r_{табл}$, то влияние фактора на признак достоверно; если меньше табличного – не достоверно, не доказано.

При использовании критерия Стьюдента для доказательства достоверности r вначале рассчитывают стандартную ошибку коэффициента корреляции:

$$m_r = \sqrt{(1 - r^2)/(N_n - 2)}, \qquad (4.3)$$

где N_n – число сопряженных пар в сравниваемых выборках.

Значение коэффициента корреляции записывают с учетом его ошибки и уровня значимости: $r_{0,95\ (0,99)} \pm m_r$. Затем вычисляют критерий Стьюдента для коэффициента корреляции:

$$t_r = r / m_r. \qquad (4.4)$$

Критерий Стьюдента можно рассчитать иначе:

$$t_r = r\sqrt{N_n - 2} / \sqrt{1 - r^2}. \qquad (4.5)$$

Если вычисленный критерий Стьюдента больше табличного (прил. 4), то зависимость существенна, если меньше – недостоверна.

Пример. Исследованиями установлено, что на длину поля севооборота влияет глубина расчленения рельефа. Необходимо доказать достоверность установленной зависимости. Получены следующие исходные данные (x – глубина расчленения рельефа, м; y – длина поля севооборота, м):

x	83	72	69	90	90	95	95	91	75	70
y	56	42	18	84	56	107	90	58	31	48

Вначале строим график (рис. 4.3), который указывает на существование между исследуемыми показателями положительной линейной зависимости, что требует вычисления коэффициента корреляции. Для этого проводим расчет данных по таблице исходных данных (4.1). Необходимые суммарные результаты подставляем в формулу (4.1) и вычисляем коэффициент корреляции:

$$r = 2302 / \sqrt{1000 \cdot 6854} = 0{,}88.$$

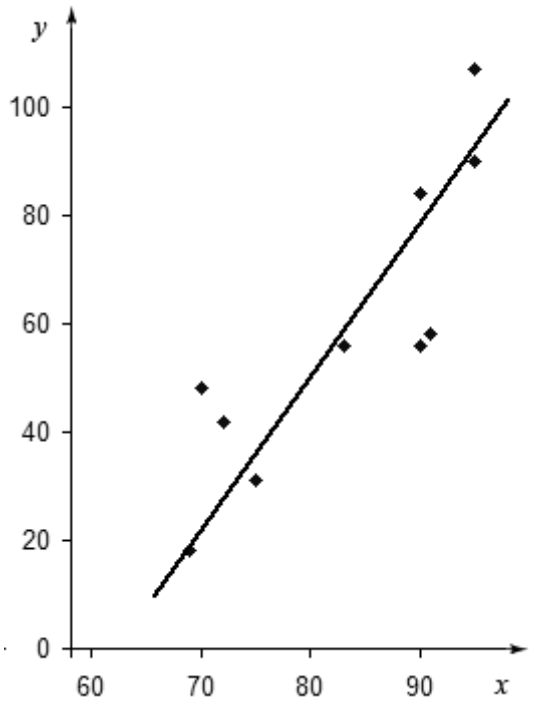


Рис. 4.3. Зависимость содержания подвижного марганца (y) от гидролитической кислотности (x)

Таблица 4.1

Исходные данные для расчета коэффициента корреляции

x_i	$x_i - M_x$	$(x_i - M_x)^2$	y_i	$y_i - M_y$	$(y_i - M_y)^2$	$(x_i - M_x) \cdot (y_i - M_y)$
69	–14	196	18	–42	1764	558
70	–13	169	48	–12	144	156
72	–11	121	42	–18	324	198
75	–8	64	31	–29	841	232
83	0	0	56	–4	16	0
90	7	49	84	24	576	168
90	7	49	56	–4	16	–28
91	8	64	68	8	64	64
95	12	144	90	30	900	360
95	12	144	107	47	2209	564
∑830 M_x 83	∑ 0	∑ 1000	∑600 M_y60	∑ 0	∑ 6854	∑ 2302

Поскольку $r_{выч} = 0{,}88 > r_{табл} = 0{,}77$ при $P = 0{,}99$ и $\nu = 8$, то зависимость между длиной поля севооборота и глубиной расчленения рельефа достоверная и положительная.

Определим также достоверность зависимости с использованием критерия Стьюдента t по формуле (4.5): $t_r = 0{,}88 \cdot \sqrt{10-2} \ / \ \sqrt{1-0{,}88^2} = 5{,}27$.

Поскольку $t_{выч} = 5{,}27 > t_{табл} = 3{,}36$ при $\nu = 8$ и $P = 0{,}99$ (см. прил. 4), то зависимость между данными показателями доказана (достоверна).

В рассмотренном примере оба критерия подтвердили достоверную линейную положительную зависимость между содержанием длиной поля севооборота и глубиной расчленения рельефа.

Таким образом, достоверность связи устанавливается путем сравнения r (η) расчетного (фактического) и r (η) теоретического (табличного). Если $r(\eta)_{выч} > r(\eta)_{табл}$ при учете степени свободы (ν) вариационных рядов и уровня вероятности P = 0,95 и 0,99, то зависимость между признаками доказана не зависимо от величины r (η). Регрессионный анализ обычно является продолжением корреляционного в случае величины $r(\eta) \geq \pm 0{,}7$.

Коэффициент детерминации R^2 (D^2) – это коэффициент корреляции, возведенный в квадрат, например, $R^2 = r^2 = 0{,}2^2 = 0{,}04$. С помощью коэффициента детерминации устанавливается доля влияния анализируемого факторного признака на результативный признак. В случае, когда $R^2 = 0{,}04$, можно утверждать, что доля влияющего фактора (x) на признак (y) составляет 4 %. Следовательно, на долю других факторов приходится 96 % влияния.

4.2. НЕЛИНЕЙНАЯ КОРРЕЛЯЦИЯ

Зависимость между признаками не всегда графически выражается в виде прямой линии. Если рассеяние точек на графике приближается к кривой линии (см. рис. 4.1, *в, г*), то зависимость устанавливается с использованием корреляционного отношения (η), величина которого изменяется только от 0 до 1. Для него теоретические значения приводятся отдельно в таблице или путем перерасчета его в критерий Стьюдента. При нелинейной корреляции вычисляется корреляционное отношение (η).

Для установления формы связи иногда используется *критерий криволинейности* в случаях, когда кривая линия мало отличается от прямой. Существует несколько способов оценки степени криволинейности. Рассмотрим два из них.

Первый способ менее точный. Оценка степени криволинейности определяется по разности коэффициента корреляции и корреляционного отношения, используя неравенство: $\eta^2 - r^2 \geq 0{,}1$. Корреляция считается криволинейной, если полученный результат соответствует этому неравенству. Предварительно следует рассчитать между сравниваемыми выборками r и η.

Второй способ оценки степени криволинейности связан с применением критерия Стьюдента:

$$t = 0{,}5 \sqrt{\frac{N}{(\eta^2 - r^2)^{-1} - 2 + (\eta^2 + r^2)}} \geq 3.$$

Если $t_{выч} < 3$ или $t_{выч} < t_{табл}$, то рассматриваемая связь несущественно отклоняется от прямолинейной, поэтому относим ее к линейной. В других случаях связь между признаками относят к криволинейной, поэтому рассчитывается корреляционное отношение.

Корреляционное отношение, как и коэффициент корреляции, используется для оценки прямой и обратной зависимости между признаками.

Оценка прямой нелинейной зависимости между признаками

Нелинейная прямая зависимость определяется как параболическая. Расчет корреляционного отношения производится по формуле с использованием функции y:

$$\eta_y = \sqrt{\frac{n\sum(\bar{y} - M_y)^2}{\sum(y_i - M_y)^2}}, \qquad (4.6)$$

где $\bar{y}$ – среднее арифметическое частных групп по y_i; n – число вариант в частной группе; $\bar{y} - M_y$ – отклонение общего среднего (M_y) от средних арифметических частных групп ($\bar{y}$).

Ошибка корреляционного отношения независимо от способа расчета вычисляется следующим образом:

$$m_\eta = \sqrt{(1-\eta^2)/(N_{\text{пар}}-2)}. \qquad (4.7)$$

Критерий Стьюдента определяется с использованием η:

$$t_\eta = \eta / m_\eta. \qquad (4.8)$$

Если $t_{\text{выч}} > t_{\text{табл}}$, то корреляционное отношение признается достоверным.

Пример. Следует установить, существует ли зависимость между температурой воздуха (x, °С) и биомассой улучшенного луга (у, т / га) по шести фермерским хозяйствам Беларуси, исходя из следующих данных:

x_i	14,7	14,9	15,3	15,6	16,0	16,7
y_i	13,3	13,7	14,2	14,5	14,7	14,6

При построении графика получена кривая, близкая к параболе (рис. 4.4).

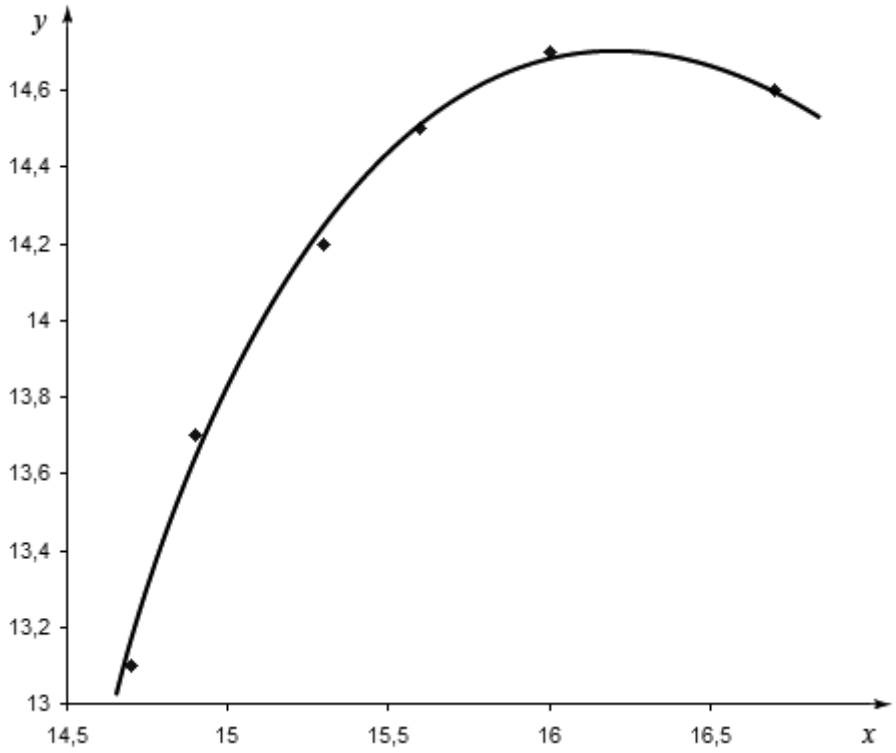


Рис. 4.4. Кривая зависимости биомассы трав (y) от температуры воздуха (x)

По исходным данным (табл. 4.2) рассчитываем корреляционное отношение между x и y. Выборку разбиваем на частные группы по значениям y. Их должно быть не менее трех. В нашем примере выделены частные две группы для сокращения расчета. Для частных групп рассчиты-

ваются средние ($\bar{y}$) и отклонение их от общей средней для выборки (M_y), а также отклонения индивидуальных вариант выборки (y_i) от общей средней (M_y). Сумму отклонений в квадрате из табл. 4.2 заносим в формулу (4.6) и вычисляем η.

$$\eta_y = \sqrt{(3 \cdot 0,40)/1,92} = 0,79.$$

Таблица 4.2

Исходные данные по биомассе трав, т / га

y_i	$\sum y_i$ по группам	$\bar{y}$, среднее по группам	$\bar{y} - M_y$	$(\bar{y} - M_y)^2$	$y_i - M_y$	$(y_i - M_y)^2$
I группа						
13,1					–1,03	1,06
13,7	41,0	13,7	–0,43	0,18	–0,43	0,18
14,2					0,07	0,005
II группа						
14,5					0,37	0,14
14,7	43,8	14,6	0,47	0,22	0,57	0,32
14,6					0,47	0,22
∑ 84,8 M_y14,13			∑ 0,04	∑ 0,40	∑ 0,02	∑ 1,92

Ошибку корреляционного отношения находим по формуле (4.7):

$$m_\eta = \sqrt{(1-(0,78)^2)/(6-2)} = 0,31.$$

Достоверность результатов определяем по критерию Стьюдента (4.8):

$t_\eta = 0,78 / 0,31 = 2,51$.

Поскольку $t_\eta = 2,51 < t_{табл} = 2,78$ при $P0,95$ для $\nu = 4$ (см. прил. 4), то значение корреляционного отношения следует признать не доказанным, а зависимость между температурой воздуха и биомассой трав положительная, но не достоверная.

Оценка обратной нелинейной зависимости между признаками

Алгоритм вычисления и доказательств при расчете корреляционного отношения обратной нелинейной (гиперболической) зависимости аналогичен алгоритму прямой нелинейной зависимости. Различие состоит в том, что в качестве исходных вариант используется выборка со значениями *x*.

Для нелинейной обратной (гиперболической) зависимости корреляционное отношение определяется с использованием аргумента *x* по формуле (4.9), условные обозначения в которой аналогичны формуле (4.6):

$$\eta_x = \sqrt{\frac{n\sum(\overline{x}_{гр} - M_y)^2}{\sum(x_i - M_x)^2}}. \qquad (4.9)$$

Расчет производится по влияющему фактору (x_i), предварительно составив таблицу по форме и получив необходимые суммы:

x_i	Σx_i по группам	$\overline{x}_{гр}$	$\overline{x}_{гр} - M_x$	$(\overline{x}_{гр} - M_x)^2$	$x_i - M_x$	$(x_i - M_x)^2$

Расчетные величины η по x сравнивают с табличными (теоретическими) для степени свободы ($v = N_{пар} - 1$) и $P = 0{,}95$ и $0{,}99$. Если расчетная величина больше табличной, то можно утверждать с уверенностью о наличии достоверной зависимости между признаком и фактором.

Для всех коэффициентов можно рассчитать их ошибки: $r \pm m_r$; $\eta \pm m_n$ и т. д.

При расчете η с использованием выборочных вариант x и y можно также применить следующие формулы с известными обозначениями:

$$\eta_x = \sqrt{\frac{\sum(x_i - M_x)^2 - (x_i - \overline{x}_{гр})^2}{\sum(x_i - M_x)^2}}, \qquad (4.10)$$

$$\eta_y = \sqrt{\frac{\sum(y_i - M_y)^2 - (y_i - \overline{y}_{гр})^2}{\sum(y_i - M_y)^2}}. \qquad (4.11)$$

4.3. ЧАСТНАЯ (ПАРЦИАЛЬНАЯ) КОРРЕЛЯЦИЯ

В практических целях часто приходится выявлять взаимодействие нескольких факторов. Производится комбинационная группировка собранного материала, которая требует большого числа наблюдений. Можно использовать специальные статистические методы. С помощью этих методов производится последовательная элиминация влияния одних факторов и выделение результатов влияния других факторов. К таким методам относится метод частной корреляции. Элиминация – это исключение неизвестного из системы уравнений.

В ходе вычисления коэффициентов частной корреляции для трех признаков последовательно элиминируется влияние одного из признаков: x_3, x_2, x_1. Элиминирование влияния третьего признака (x_3) и выявление связи между x_1 и x_2 производится по формуле:

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - {r_{13}}^2)(1 - {r_{23}}^2)}}. \quad (4.12)$$

Аналогично производится элиминирование влияния второго признака (x_2) и выявление связи между x_1 и x_3:

$$r_{13,2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - {r_{12}}^2)(1 - {r_{23}}^2)}}. \quad (4.13)$$

Затем проводится элиминирование влияния первого признака (x_1) и выявление взаимосвязи x_2 и x_3:

$$r_{23,1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1 - {r_{13}}^2)(1 - {r_{12}}^2)}}. \quad (4.14)$$

Пример. Оценить взаимосвязь фактора длительности рабочего времени механизатора на уборке картофеля усталостью (число оставленных кустов – ошибок) и производительностью труда (собранной массой клубней картофеля) (табл. 4.3).

Рассчитав коэффициенты корреляции Пирсона r_{12} = 0,4; r_{13} = –0,7; r_{23} = –0,4, можно сделать выводы о влиянии на появление усталости длительности рабочего времени (r_{12}) и снижении производительности труда с увеличением продолжительности работы (r_{13}). Между увеличением усталости и снижением производительности труда обнаружена обратная статистическая связь (r_{23}).

Таблица 4.3

Исходные данные для расчета коэффициентов частной корреляции

Время работы, ч	Число ошибок	Масса клубней
4	5	4
1	6	6
3	6	2
3	6	6
5	6	4
2	7	3
1	8	1
5	8	3
6	9	1
6	9	1

Используя формулы коэффициентов частной корреляции, произведем их расчет:

$$r_{12,3} = \frac{0,4 - (-0,7)(-0,4)}{\sqrt{(1-(-0,7)^2)(1-(-0,4)^2)}} = 0,2,$$

$$r_{23,1} = \frac{-0,4 - 0,4(-0,7)}{\sqrt{(1-(-0,7)^2)(1-0,4^2)}} = -0,1,$$

$$r_{13,2} = \frac{-0,7 - 0,4(-0,4)}{\sqrt{(1-(-0,4)^2)(1-0,4^2)}} = -0,7.$$

Таблица 4.4

Итоговые значения коэффициентов корреляции

r_{12}	0,4	$r_{12,3}$	0,2
r_{13}	–0,7	$r_{23,1}$	–0,1
r_{23}	–0,4	$r_{13,2}$	–0,7

Анализ коэффициентов в табл. 4.4 показывает, что при устранении фактора продолжительности труда, произошел сдвиг показателя $r_{23,1}$ = –0,1 (связь между усталостью и производительностью труда исчезла). Снижение выработки продукции к концу рабочего дня связано в первую очередь не с нарастанием усталости, а с какими-то другими причинами.

4.4. ПОНЯТИЕ О МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ

Метод множественной корреляции применяется в случаях, когда необходимо установить совокупное влияние всего комплекса факторов на результативный признак. Величина коэффициента множественной корреляции изменяется от 0 до 1. Его можно вычислить с использованием коэффициентов частной линейной корреляции по формуле:

$$R_{1,23} = \sqrt{1-(1-r_{12}{}^2)(1-r_{13}{}^2)} = \sqrt{1-(0,4^2)(1-(-0,7)^2)} = 0,75.$$

По коэффициенту R = 0,75 определяется коэффициент детерминации R^2 (R_D) = $0,75^2$ = 0,56. Он показывает, что доля совместного влияния второго и третьего признаков составляет 56 %.

Таким способом можно устанавливать влияние других признаков на второй или третий признак. Тем самым дается конкретная оценка влияния факторов на признак.

4.5. ОЦЕНКА РАЗЛИЧИЙ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ

Решение задач по оценке различий между коэффициентами корреляции возникает иногда в случае, если обе выборки принадлежали к одной генеральной совокупности.

Пример. Требуется оценить статистическую достоверность различий между коэффициентами $r_1 = 0{,}45$; $r_2 = 0{,}58$. Число наблюдений в первой и второй группах составило соответственно $N_1 = 74$ и $N_2 = 50$.

По таблице величин Z [$r = \varphi(Z)$] (прил. 8) значения коэффициентов корреляции переводятся в соответствующие им величины $Z_1 = 0{,}48$, $Z_2 = 0{,}66$.

Оценка производится по критерию Стьюдента:

$$t = | Z_2 - Z_1 | / \sqrt{(N_1 + N_2)/[(N_1 - 3)(N_2 - 3)]};$$

$$t = | 0{,}66 - 0{,}48 | / \sqrt{(74 + 50)/[(74 - 3)(50 - 3)]} = 1{,}09. \qquad (4.15)$$

Число степеней свободы равно $N_1 + N_2 - 4 = 74 + 50 - 4 = 120$. При уровне значимости $\alpha = 0{,}05$ критическая величина критерия Стьюдента составляет 1,98 (прил. 4), что больше вычисленного (1,09). Поэтому различия между r_1 и r_2 следует признать статистически недостоверными.

Следует иметь в виду, что при анализе коэффициентов корреляции: чем больше r, тем меньшие различия между ними становятся значимыми. Если для $r = 0{,}14$ и 0,24 (разница между ними в 0,1) может быть статистически не значимой, то для $r = 0{,}80$ и 0,90 (разница 0,1) может оказаться значимой.

4.6. РАНГОВАЯ КОРРЕЛЯЦИЯ

В географических исследованиях иногда приходится обрабатывать быстро и с наименьшими затратами фактический материал, даже если получаются менее точные результаты. В некоторых случаях работают с качественной информацией или с громоздкими вычислениями. В таких случаях для установления зависимости между признаками используется ранговая корреляция.

Процесс упорядочения вариант по какому-либо признаку (например, увеличение или уменьшение количества населения по районам) называют ранжированием. Каждому члену ранжированного ряда присваивается *ранг.* Для обозначения рангов, как правило, используются числа в пределах единиц и десятков, например: 1, 2, 3, ..., *n.* Первой варианте или группе вариант присваивается ранг 1, второй варианте или группе – 2 и т. д. Следует иметь в виду, что одни и те же варианты в зависимости от цели

группировки могут иметь различные ранги. Величина ранга не позволяет нам судить о том, насколько близко друг к другу расположены на шкале измерения различные варианты совокупности или качественные признаки.

Ранговую корреляцию можно применять для всех упорядоченных признаков (например, экспертные оценки, баллы, бонитеты). Объем сопряженных выборок должен быть не менее пяти. Коэффициент ранговой корреляции характеризуется следующими свойствами.

1. Если ранжированные варианты выборочных совокупностей имеют один и тот же ранг независимо от цели ранжирования, то коэффициент корреляции должен быть равен +1, т. е. существует полная положительная функциональная зависимость:

N_1	1	2	3	4	5	6	7
N_2	1	2	3	4	5	6	7

2. Если ранги вариант в сравниваемых рядах выборочных совокупностей расположены в обратной последовательности, то коэффициент корреляции равен –1, т. е. будет иметь место полная обратная функциональная зависимость:

N_1	1	2	3	4	5	6	7
N_2	7	6	5	4	3	2	1

3. В других случаях коэффициент ранговой корреляции имеет значения между +1 и –1, что больше соответствует фактической связи между признаками.

Для расчета зависимости (x, y) существуют следующие коэффициенты ранговой корреляции: коэффициент неупорядоченности $r_{\text{н}}$ и коэффициент Спирмена $r_{\text{С}}$. Коэффициент ранговой корреляции Спирмена рассчитать легче, чем коэффициент неупорядоченности, поэтому в естественных науках предпочтение отдается $r_{\text{С}}$. Коэффициент Спирмена представляет собой следующее соотношение:

$$r_C = 1 - \frac{6\sum(d^2)}{N_n{}^3 - N_n}, \text{ или } r_C = 1 - \frac{6\sum(x' - y')^2}{N_n{}^3 - N_n}, \qquad (4.16)$$

где d – разность между сопряженными рангами; x' – величины рангов, заменяющие фактические варианты или качественные признаки по аргументу x; y' – величины рангов, заменяющие фактические варианты или качественные признаки по функции y; $N_{\text{п}}$ – количество сопряженных пар.

Достоверность полученного рангового коэффициента можно установить аналогично достоверности коэффициента корреляции (прил. 9).

Пример. Следует дать эстетическую оценку ландшафта для обоснования выбора места застройки хозяйственных объектов. Предложено срав-

нить пять видов ландшафта (аргумент x), имеющих свои преимущества с точки зрения чистоты и влажности воздуха, насыщенности полезными фитонцидами, характеризующихся разнообразием рельефа, растительности, наличием рек и водоемов.

Исходя из имеющихся показателей, расположим виды ландшафта с учетом возрастающей оздоровительной и эстетической их роли (табл. 4.5). Соответственно этому видам ландшафта присваиваются ранги по возрастающей величине. Для получения необходимых показателей при расчете рангового коэффициента корреляции составляем табл. 4.6. Вычисляем разность между парными рангами ($x' - y'$), которые возводим в квадрат и суммируем. Результаты используются для расчета рангового коэффициента корреляции по формуле (4.6):

$$r_C = 1 - \left[6 \cdot 1 : (125 - 5)\right] = 0{,}95.$$

Таблица 4.5

Оценка ландшафта для хозяйственных целей

Вид ландшафта	Ранг x'	Самочувствие работников	Ранг y'
Плоский пониженный, со смешанным лесом на суглинистых почвах	1	удовлетворительное	1
Слегка волнистый, с ельником на суглинистых почвах	2	удовлетворительное	1
Всхолмленный, с сосново-лиственным лесом и водоемом на песчаных почвах	3	хорошее	3
Пересеченный, с сосновым лесом на песчаных и супесчаных почвах	3	хорошее	3
Слегка пересеченный, с сосново-можжевеловым лесом на песчаных и супесчаных почвах	4	отличное	4

Поскольку ранговый коэффициент корреляции $r_C = 0{,}95 > r_т = 0{,}80$ при $P = 0{,}90$ для $\nu = 4$ (прил. 9 Спирмена), можно сделать вывод, что влияние изучаемых ландшафтов на самочувствие работающих достоверно и положительно.

Таблица 4.6

Расчет рангового коэффициента корреляции

x'	y'	$x' - y'$	$(x' - y')^2$
1	1	0	0
2	1	1	1
3	3	0	0
3	3	0	0
4	4	0	0

Глава 5. РЕГРЕССИОННЫЙ АНАЛИЗ

Регрессионный анализ является продолжением корреляционного анализа. Он развивает и углубляет представление о корреляционной связи. Если корреляционный анализ позволяет установить лишь форму и тесноту зависимости между случайными переменными, то регрессионный анализ математически описывает выявленную зависимость, т. е. дает возможность численно оценить одни параметры через другие. Составив и решив уравнения регрессии, можно произвести выравнивание эмпирических линий регрессии, т. е. моделировать наблюдаемую зависимость путем подбора функции, график которой представляет собой теоретическую линию регрессии. Если подобранная функция отражает сущность процесса или явления, то возможно прогнозирование значений признака за пределами сделанных наблюдений. Подобно корреляции, регрессия может быть *парной* (простой) и *множественной,* по форме связи – *линейной* и *нелинейной*, по зависимости – *односторонней* (изменяется лишь один признак под влиянием другого) и *двусторонней* (изменяются оба признака под воздействием друг друга).

Регрессия выражается несколькими способами: построением эмпирических линий, составлением уравнения и затем – построением теоретических линий регрессии, а также с помощью коэффициента регрессии. Уравнение наиболее точно выражает зависимость между двумя переменными (x, y), если корреляция между ними близка к единице.

Регрессионный анализ возможен при наличии всего лишь нескольких пар сопряженных наблюдений, но при условии сильных связей между признаками ($r \geq 0{,}7$). Для вывода уравнения линейной регрессии достаточно двух пар наблюдений. Обычно рядом с уравнением регрессии приводится коэффициент корреляции или корреляционного отношения, например: $y = 0{,}1106\,x + 0{,}298$, $r_{0{,}95} = 0{,}75$ (это обусловлено практическим использованием уравнения регрессии). Из приведенных равенств вытекает, что влияние аргумента (x) на функцию (y) достаточно сильное. Поэтому, имея в своем распоряжении данные по аргументу, можно по формуле уравнения регрессии вычислить значение функции, не прибегая к полевым наблюдениям.

Точки эмпирических линий регрессии ($\bar{x}_{\text{гр}}, \bar{y}_{\text{гр}}$) определяются как взвешенные средние арифметические, для невзвешенных рядов – как средние малых групп выборки. Вычислив координаты точек, наносим их на график и соединяем прямой; в результате получаются эмпирические линии регрессии (рис. 5.1). По графическому изображению можно пред-

варительно сделать заключение о характере связи. При полном отсутствии связи эмпирические линии располагаются параллельно осям графика. При полной связи между x, y ($r = 1$) линии регрессии на графике, построенные по точкам эмпирических линий регрессии, совместятся.

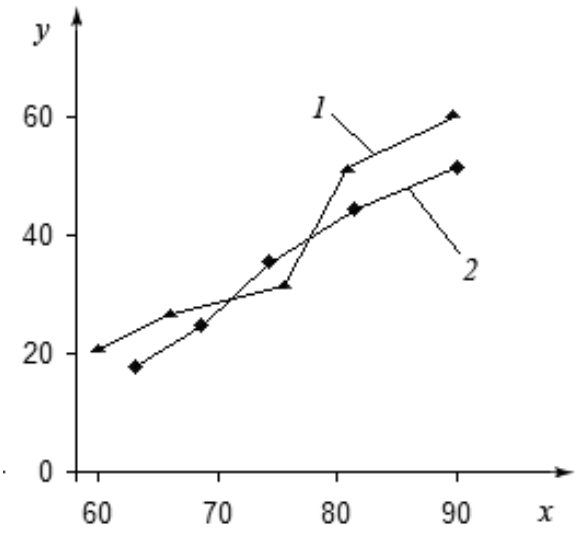


Рис. 5.1. Эмпирические линии регрессии по $\overline{x}_{гр}$, $\overline{y}_{гр}$

Существует два способа составления уравнений регрессии: а) способ координат точек, с использованием двух–трех точек, расположенных на эмпирической линии (желательно в начале, середине и конце ее), – для тех случаев, когда расчет не требует большой точности; б) способ наименьших квадратов, более точный, так как для составления уравнения регрессии привлекаются все сопряженные наблюдения. Рассмотрим наиболее простые способы составления уравнений регрессии.

5.1. ЛИНЕЙНАЯ ЗАВИСИМОСТЬ

Линейная регрессия на графике изображается в виде прямой так, чтобы точки эмпирической линии располагались по обе стороны ее и по возможности ближе к ней.

Известно следующее уравнение линейной регрессии:

$$y = ax + b, \qquad (5.1)$$

где y – значение зависимой переменной (признак); x – значение независимой переменной (фактор, влияющий на признак); a – коэффициент регрессии, показывающий степень зависимости между переменными (может

быть также выражен тангенсом угла наклона линии регрессии к оси абсцисс); b – ордината линии, показывающая смещение начала прямой относительно начала координат.

Определим двумя способами неизвестные параметры a и b. Используем для этого пример нахождения линейной корреляции (см. п. 4.1).

Пример. Следует установить, как влияет гидролитическая кислотность (x_i, мэкв. на 100 г почвы) на содержание подвижного марганца (y_i, мг/кг почвы). В результате аналитических работ получены следующие данные:

x_i	69	70	72	75	83	90	91	95	95
y_i	18	48	42	31	56	84	68	90	107

Для решения поставленной задачи используем *способ координат точек*. Результаты наблюдений наносим на график, затем проводим прямую так, чтобы число точек по обе стороны линии было одинаковым (рис. 5.2). Для расчета параметров a и b выбираем две точки, которые находятся на прямой или рядом с ней (одну в начале и одну в конце). Используем координаты точек 1-й и 8-й: $x_1 = 69$, $y_1 = 18$; $x_8 = 95$, $y_8 = 90$. Подставляя значения переменных в общее уравнение прямой, получаем систему уравнений:

$$\begin{cases} 18 = 69a + b; \\ 90 = 95a + b. \end{cases}$$

Решаем эту систему относительно a и b: $b = 18 - 69\,a$; $90 = 95a + (18 - 69a)$; $72 = 26a$; $a = 2{,}76$ (или $\mathrm{tg} = 70^{\circ}06'$); $b = 18 - 69 \cdot 2{,}76 = -173{,}07$. Получив количественное значение параметров a и b, связь между x и y можно выразить конкретным уравнением регрессии:

$$y = 2{,}76x - 173{,}07,\ r_{0{,}99} = 0{,}87.$$

Это уравнение можно использовать для расчета содержания марганца, если имеются данные по гидролитической кислотности (с учетом заданных условий).

Приведенное выше уравнение регрессии можно получить также способом наименьших квадратов, используя координаты всех точек. Этот способ заключается в построении такой линии на графике, чтобы сумма квадратов отклонений от нее до точек эмпирической линии регрессии была наименьшей. Для определения параметров a и b составляется система уравнений:

$$\begin{cases} \Sigma y = a\Sigma x + bn; \\ \Sigma xy = a\Sigma x^2 + b\Sigma x. \end{cases} \qquad (5.2)$$

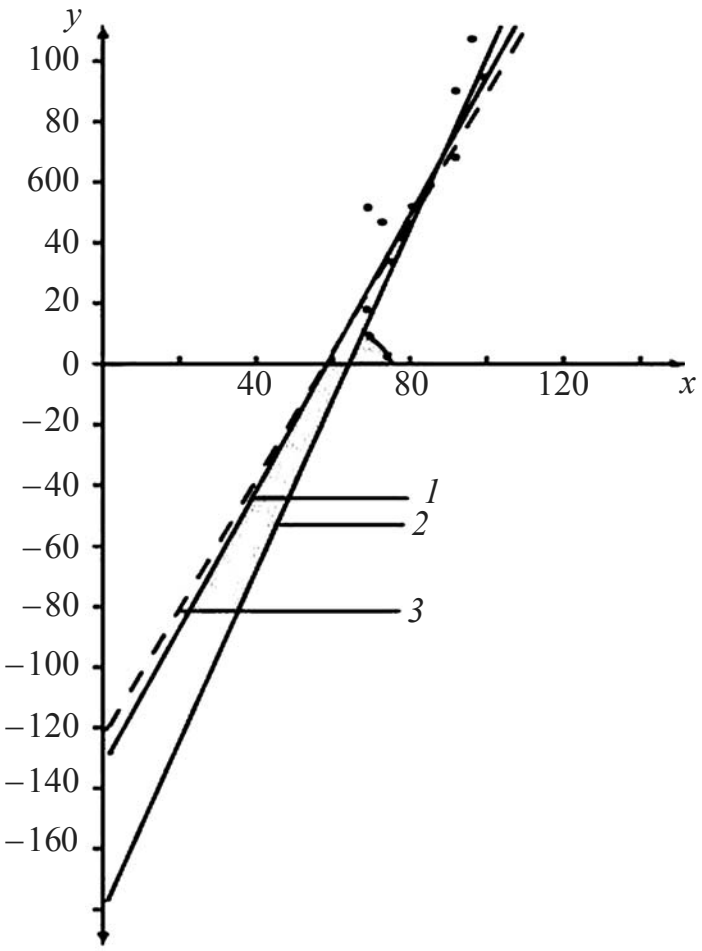


Рис. 5.2. Сравнение местоположения
эмпирических линий (*1*, *2*) с теоретической (*3*)
по зависимости содержания подвижного марганца *y*
от гидролитической кислотности *x* ($\angle a = 70^{o}06' = \mathrm{tg}_x\ 2{,}76$):
для эмпирической линии 1 $y = 2{,}30x - 130{,}9$;
для 2 $y = 2{,}76x - 173{,}0$; $r_{0{,}99} = 0{,}87$

Систему уравнений выводим следующим образом. Подставляем в общее уравнение прямой (5.1) все имеющиеся значения по гидролитической кислотности (x) и содержанию подвижного марганца (y), суммируем правые и левые части и получаем первое уравнение:

$$\begin{aligned} y_1 &= ax_1 + b; \\ y_2 &= ax_2 + b; \\ &\cdots\cdots\cdots\cdots \\ \underline{y_n &= ax_n + b} \\ \Sigma y &= a\Sigma x + bn. \end{aligned} \qquad (5.3)$$

Затем каждое исходное уравнение из (5.3) умножаем на соответствующее значение *x;* просуммировав правые и левые части, получим второе уравнение:

$$x_1 y_1 = a x_1^2 + b x_1;$$
$$x_2 y_2 = a x_2^2 + b x_2;$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$\underline{x_n y_n = a x_n^2 + b x_n}$$
$$\Sigma xy = a\Sigma x^2 + b\Sigma x.$$

Для расчета параметров a и b составляем табл. 5.1. Полученные данные подставляем в систему уравнений (5.2):

$$\begin{cases} 600 = 830a + 10b; \\ 52102 = 69890a + 830b. \end{cases}$$

Таблица 5.1

Расчет данных для уравнения линейной зависимости

x	y	xy	x^2	$y' = ax + b$	Расчет критерия χ^2		
					$y - y'$	$(y - y')^2$	$\frac{(y - y')^2}{y'}$
69	18	1242	4761	27,8	–9,8	96,04	3,45
70	48	3360	4900	30,1	17,9	320,41	10,64
72	42	3024	5184	34,7	7,3	53,29	1,54
75	31	2325	5625	41,6	–10,6	112,36	2,70
83	56	4648	6889	60,0	–4,0	16,00	0,27
90	84	7560	8100	76,1	7,9	62,41	0,82
90	56	5040	8100	76,1	–20,1	404,01	5,31
91	68	6188	8281	78,4	–10,4	108,16	1,38
95	90	8550	9025	87,6	2,4	5,76	0,07
95	107	10 165	9025	87,6	19,4	376,36	4,30
Σ 830	600	52 102	69 860				$\chi^2 = 30,47$

Решая систему, находим искомые параметры: a = 2,30 (tg = 66°30′); b = –130,9. Подставив полученные показатели в искомое уравнение регрессии, находим $y = 2,30\,x - 130,9$; $r_{0,99} = 0,87$.

Хотя значения параметров a и b, рассчитанные двумя способами, близки между собой, второй способ (наименьших квадратов) более точно определяет положение линии регрессии.

Кроме того, коэффициенты *a* и *b* для уравнения регрессии также могут быть рассчитаны на основе исходных данных (*x*, *y*) по формулам, которые обеспечивают наименьший квадрат отклонений этих точек от линии регрессии (метод наименьших квадратов):

$$b = \frac{\Sigma x^2 \Sigma y^2 - \Sigma x \Sigma xy}{n_{\text{пар}} \Sigma x^2 - (\Sigma x)^2}, \quad a = \frac{n_{\text{пар}} \Sigma xy - \Sigma x \Sigma y}{\Sigma x^2 - (\Sigma x)^2}.$$

После составления уравнения регрессии и определения параметров *a* и *b* производим расчет точек *y′* теоретической линии регрессии. Для этого в уравнение регрессии поочередно подставляем значения *x*.

Степень совпадения теоретической и эмпирической линии регрессии можно проверить, используя критерий хи-квадрат. Цифровые показатели для $(y - y')^2 / y'$ (см. табл. 5.1) суммируем и получаем $\chi^2 = 30{,}47$. Поскольку $\chi_ф^2 = 30{,}47 > \chi_т^2 = 21{,}66$ при $P = 0{,}99$ для $\nu = 9$, то можно указать на недостаточное соответствие теоретической линии регрессии эмпирическому ряду. Составленные уравнения регрессии можно проверить на точность зависимости между переменными *(x, y)* не только по критерию хи-квадрат, но и по коэффициенту точности выравнивания линии r_1, отражающему степень приближения (соответствия) фактических данных наблюдения к вероятным. Этот коэффициент определяем следующим образом:

$$r_1 = \sqrt{\frac{\Sigma\alpha^2 - \Sigma\beta^2}{\Sigma\alpha^2}} = \sqrt{\frac{\Sigma(y_ф - M_ф)^2 - \Sigma(y_ф - y_в)^2}{\Sigma(y_ф - M_ф)^2}}, \qquad (5.4)$$

где $(y_ф - M_ф) = \alpha$ – отклонение индивидуальных вариант от общего среднего арифметического по *y;* $(y_ф - y_в) = \beta$ – отклонение индивидуальных экспериментальных вариант по *y* от расчетных по уравнению.

На основании исходных данных, полученных в табл. 5.2, используя формулу (5.4), имеем:

$$r_1 = \sqrt{(6806 - 1554{,}8) : 6806} = 0{,}88.$$

Принято считать: если $r_1 > 0{,}95$, то уравнение регрессии соответствует более точному положению линии на графике. При $r_1 < 0{,}95$ необходимо найти другую математическую зависимость. В приведенном примере $r_1 = 0{,}88 < 0{,}95$, поэтому следует подобрать другую математическую зависимость. Такие же выводы получены при проверке на точность зависимости между переменными по критерию хи-квадрат. Оба критерия оценки (χ^2, r_1) на точность выравнивания линии уравнения регрессии используются и для других форм регрессионной зависимости.

Таблица 5.2

Расчет данных для определения точности выравнивания линии

y		α-отклонения		β-отклонения	
$y_ф$	$y_в$	$y_ф - M_ф$	$(y_ф - M_ф)^2$	$y_ф - y_в$	$(y_ф - y_в)^2$
18	27,8	–42	1764	–9,8	96,04
48	30,1	–12	144	17,9	320,41
42	34,7	–18	324	7,3	53,29
31	41,6	29	841	–10,6	112,36
56	60,0	–4	16	–4,0	16,00
84	76,1	24	576	7,9	62,41
56	76,1	–4	16	–20,1	404,01
68	78,4	4	16	–10,4	108,16
90	87,6	30	900	2,4	5,76
107	87,6	47	2209	19,4	376,36
$M_ф = 60$			Σ 6806		Σ 1554,80

Ошибку уравнения регрессии можно определить по формуле

$$m = \sqrt{\frac{\Sigma(y - y')^2}{n - k}} = \sqrt{\frac{\Sigma(y_ф - y_в)^2}{n - k}},$$

где n – число точек линии регрессии (см. рис. 5.2); k – число коэффициентов в уравнении регрессии (два плюс свободный член уравнения).

5.2. ГИПЕРБОЛИЧЕСКАЯ ЗАВИСИМОСТЬ

При проведении исследований может быть установлена нелинейная зависимость между аргументом и функцией, представляющая собой на графике кривую в виде гиперболы. Общее уравнение регрессии для гиперболической зависимости имеет вид

$$y = a / x + b, \qquad (5.5)$$

где x – аргумент; y – функция; a и b – коэффициенты, величину которых следует установить.

Расчет сводится к следующему. Чтобы установить вид зависимости между функцией и аргументом, по исходным данным строится график. Затем при вычислении параметров a и b по способу координат точек подбираются две точки, расположенные на кривой или около нее по методу, описанному для линейной регрессии (см. п. 5.1). Для этих же параметров по способу наименьших квадратов используется система уравнений

$$\begin{cases} \Sigma xy = an + b\Sigma x; \\ \Sigma x^2 y = a\Sigma x + b\Sigma x^2. \end{cases} \qquad (5.6)$$

Эта система получена путем умножения на x и x^2 исходных уравнений по x и y:

$$\begin{array}{ll} x_1 y_1 = a + bx_1; & x_1^2 y_1 = ax_1 + bx_1^2; \\ x_2 y_2 = a + bx_2; & x_2^2 y_2 = ax_2 + bx_2^2; \\ \ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots\ldots \\ \underline{x_n y_n = a + b\Sigma x_n} & \underline{x_n^2 y_n = ax_n + bx_n^2} \\ \Sigma xy = an + b\Sigma x. & \Sigma x^2 y = a\Sigma x + b\Sigma x^2. \end{array}$$

Пример. Установим зависимость между температурой воздуха в июле (x, °С) и относительной влажностью воздуха (y, %) по следующим исходным данным:

x_i	14,7	14,9	15,3	15,6	16,0	16,7
y_i	80	78	76	75	74	73,7

При построении графика видно, что зависимость между функцией и аргументом гиперболическая, поэтому используем общее уравнение гиперболы. Для расчета параметров a и b по способу координат точек используем данные первой и шестой пары наблюдений: $x_1 = 14{,}7$, $y_1 = 80$; $x_в = 16{,}7$, $y_в = 73{,}7$. Подставляем эти данные в общее уравнение (5.5), предварительно преобразовав его: $xy = a + bx$. Получим систему уравнений

$$\begin{cases} 1176 = a + 14{,}7b; \\ 1230{,}8 = a + 16{,}7b. \end{cases}$$

Решаем систему относительно a и b: $a = 773{,}22$; $b = 27{,}4$. В результате конкретное уравнение регрессии для гиперболической зависимости по способу координат точек будет иметь вид $y = 773{,}22 / x + 27{,}4$; $\eta_{0,99} = 0{,}84$.

Для установления параметров a и b по способу наименьших квадратов по уравнению (5.6) предварительно проводим соответствующие вычисления (табл. 5.3). Полученные данные подставляем в уравнение (5.6):

$$\begin{cases} 7086 = 6a + 93{,}2b; \\ 22\,615{,}9 = 93{,}2a + 1450{,}44b. \end{cases}$$

Делим первое уравнение на 6, второе уравнение – на 93,2 и освобождаемся от коэффициентов при неизвестном a. Затем вычитаем второе уравнение из первого и определяем b. Подставив значение b в одно из уравнений, вычисляем a. Искомое уравнение регрессии примет вид

$$y = 484\,597{,}4 / x - 3\,1280; \ \eta_{0,95} = 0{,}84.$$

Таблица 5.3

Расчет данных для уравнения гиперболической зависимости

x	y	xy	x^2	x^2y
14,7	80,0	1176,0	216,09	17 287,2
14,9	78,0	1162,2	222,01	17 316,78
15,3	76,0	1162,8	234,09	17 790,84
15,6	75,0	1170,0	243,36	18 252,00
16,0	74,0	1184,0	256,00	18 994,00
16,7	73,7	1230,79	278,89	20 554,19
Σ93,2	456,7	7085,79	1450,44	110 195

Коэффициент точности выравнивания линии r_1 по формуле (5.4) рассчитываем таким же образом, как в п. 5.1.

5.3. ПАРАБОЛИЧЕСКАЯ ЗАВИСИМОСТЬ

Общее уравнение параболы n-го порядка имеет вид

$$y = ax^n + bx^{n-1} + cx^{n-2} + \ldots + kx + 1.$$

Если ограничиться второй ступенью независимой переменной величины x, будем иметь частный случай параболы второго порядка:

$$y = ax^2 + bx + c. \qquad (5.7)$$

Пример. Для решения конкретной задачи используем данные примера из п. 4.2, где было рассчитано прямое корреляционное отношение ($\eta = 0{,}78$) и доказана его достоверность. На графике зависимость между температурой воздуха (x) и упругостью водяного пара (y) имеет вид параболы.

Для расчета коэффициентов a, b, c способом координат точек используем координаты 2-й, 4-й и 6-й точек:

$$x_2 = 14{,}9; \quad x_4 = 15{,}6; \quad x_6 = 16{,}7;$$
$$y_2 = 13{,}7; \quad y_4 = 14{,}5; \quad y_6 = 14{,}6.$$

Подставляя значения координат точек в общее уравнение параболы второго порядка (5.7), получаем систему уравнений, которую решаем относительно a, b, c:

$$\begin{cases} 13{,}7 = 222{,}01a + 14{,}9b + c; \\ 14{,}5 = 243{,}36a + 15{,}6b + c; \\ 14{,}6 = 278{,}89a + 16{,}7b + c. \end{cases}$$

В результате $a = 0{,}066$; $b = -0{,}19$; $c = 1{,}94$. С помощью уравнения параболы (5.7) имеем следующую зависимость между переменными: $y = 0{,}066x^2 - 0{,}19x + 1{,}94$, $\eta_{0,95} = 0{,}78$.

Для вычисления коэффициентов a, b, c по способу наименьших квадратов используется общее уравнение параболы второго порядка. Подставив в формулу (5.7) все имеющиеся данные и просуммировав правые и левые части уравнений, получаем первое уравнение системы:

$$y_1 = ax_1^2 + bx_1 + c;$$
$$y_2 = ax_2^2 + bx_2 + c;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\underline{y_n = ax_n^2 + bx_n + c;}$$
$$\Sigma y = a\Sigma x^2 + b\Sigma x + cn.$$

Второе и третье уравнения системы определяем путем умножения соответственно на x и x^2 исходного общего уравнения параболы второго порядка. В результате имеем систему трех уравнений:

$$\begin{cases} \Sigma y = a\Sigma x^2 + b\Sigma x + cn; \\ \Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x; \\ \Sigma x^2 y = a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2. \end{cases} \qquad (5.8)$$

Конкретные данные для уравнения (5.8) рассчитаны по табл. 5.4.

Таблица 5.4

Расчет данных для уравнения параболической зависимости

x	y	xy	x^2	x^3	x^2y	x^4
14,7	13,1	192,57	216,09	3176,52	2830,78	46 694,89
14,9	13,7	204,13	222,01	3307,95	3041,54	49 288,44
15,3	14,2	217,26	234,09	3581,58	3324,08	54 798,13
15,6	14,5	226,20	243,36	3796,42	3528,72	59 224,09
16,0	14,7	235,20	256,00	4096,00	3763,20	65 536,00
16,7	14,6	243,82	278,89	4657,46	4071,79	77 779,63
Σ 93,2	84,8	1319,18	1450,44	22 615,93	20 560,10	353 321,18

Пример. Решаем систему (5.8) относительно a, b, c:

$$84{,}8 = 1450{,}44a + 93{,}2b + 6c;$$
$$1319{,}18 = 22\,615{,}93a + 1450{,}44b + 93{,}2c;$$
$$20\,560{,}10 = 353\,321{,}18a + 22\,615{,}93b + 1450{,}44c.$$

Имеем параметры a, b, c: $a = -0{,}014$; $b = 1{,}13$; $c = -0{,}93$. Таким образом, уравнение параболы 2-го порядка, полученное по способу наименьших квадратов, примет следующий вид: $y = -0{,}014x^2 + 1{,}13x - 0{,}093$.

Сравним уравнения параболы, полученные двумя способами, подставив в эти уравнения одно из значений x (14,7 °С):

$$y = 0{,}066x^2 - 0{,}19x + 1{,}94 = 14{,}26 - 2{,}79 + 1{,}94 = 13{,}41 -$$

по способу координат точек;

$$y = -0{,}014x^2 + 1{,}13x - 0{,}093 = 3{,}02 + 16{,}61 - 0{,}093 = 13{,}49 -$$

по способу наименьших квадратов.

5.4. МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

При установлении зависимости между признаками иногда используется больше одной независимой переменной. В таком случае применяют *множественный регрессионный анализ.* Проведение анализа возможно в следующих условиях: распределение зависимой переменной при различных значениях независимых должно быть близко к нормальному; дисперсия зависимой переменной при разных значениях признаков x должна считаться одинаковой. С увеличением числа признаков и в случаях нелинейной множественной регрессии необходимо использовать ЭВМ. Поэтому рассмотрим простой вариант множественной линейной регрессии без применения ЭВМ, когда один признак зависит от двух факторов. Общее уравнение линейной множественной регрессии имеет вид:

$$y = a + bx + cz. \tag{5.9}$$

Для вычисления параметров a, b, c составляется следующая система уравнений:

$$\begin{cases} \Sigma y = an + b\Sigma x + c\Sigma z; \\ \Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma xz; \\ \Sigma yz = a\Sigma z + b\Sigma xz + c\Sigma z^2. \end{cases} \tag{5.10}$$

Соответствие между теоретическими (y') и эмпирическими (y) значениями признака устанавливают с помощью критериев хи-квадрат или Стьюдента (см. п. 1.6). При необходимости ошибка уравнения линейной множественной регрессии определяется по формуле:

$$m = \sqrt{\frac{\Sigma a_y^2 - (b\Sigma a_y a_x + c\Sigma a_y a_z)}{n-k}}, \qquad (5.11)$$

где a, b, c – значения параметров уравнения множественной регрессии; n – число сопряженных значений вариант; k – число коэффициентов уравнения регрессии (a, b, c плюс свободный член).

Другие параметры для (5.11) вычисляются по формулам:

$$\Sigma a_y^2 = \Sigma y^2 - n M_y^2; \qquad (5.12)$$

$$\Sigma a_y a_x = \Sigma xy - n M_y M_x; \qquad (5.13)$$

$$\Sigma a_y a_z = \Sigma yz - n M_y M_z. \qquad (5.14)$$

Пример. При изучении зависимости между биомассой трав (y, г/м^2) в агроландшафте, с одной стороны, температурой (x,°С) и количеством атмосферных осадков (z, мм) с другой, установлена прямая односторонняя зависимость y от x и z. С практической точки зрения целесообразно составить уравнение множественной регрессии, которое можно было бы использовать для прогноза биомассы по температуре и количеству выпавших осадков. Данные по x, y, z представляют собой средние многолетние показатели за период вегетации (май, июнь):

y	300	350	370	420	450	500
x	14,5	15,0	15,6	17,2	18,5	19,3
z	82	95	105	120	130	140

Вычисленные в табл. 5.5 показатели подставляем в систему уравнений (5.10):

$$\begin{cases} 2390 = 6a + 100,09b + 672c; \\ 405\ 71 = 100,09a + 1689,19b + 11\ 423c; \\ 275\ 600 = 672a + 11\ 423b + 77674c. \end{cases}$$

Таблица 5.5

Расчет для уравнения линейной множественной регрессии

y	x	z	y^2	x^2	z^2	xy	xz	yz
300	14,5	82	90 000	210,25	6724	4350	1189	24 600
350	15,0	95	122 500	225,00	9025	5250	1425	33 250
370	15,6	105	136 900	243,36	11 025	5772	1638	38 850
420	17,2	120	176 400	295,84	14 400	7224	2064	50 400
450	18,5	130	202 500	342,25	16 900	8325	2405	58 500
500	19,3	140	250 000	372,49	19 600	9650	2702	70 000
Σ 2390	100,1	672	978 300	1689,19	77 674	40 571	11423	275 600
$M_y = 398,33$	$M_x = 16,68$	$M_z = 112$						

Решаем систему уравнений относительно a, b, c. Получаем $a = –3{,}26$; $b = 5{,}01$; $c = 2{,}84$. Подставляем значения a, b, c в общую формулу уравнения множественной регрессии (5.9):

$$y = –3{,}26 + 5{,}01x + 2{,}84z. \qquad (5.15)$$

Затем находим теоретические значения y'. Для этого подставляем в формулу (5.15) экспериментальные данные по x и z и заносим в табл. 5.6 для расчета критерия хи-квадрат. Поскольку $\chi^2_ф = 0{,}602 < \chi^2_т = 11{,}1$ при $P = 0{,}95$ для $\nu = 5$, то можно сделать вывод, что расчет биомассы по данным температуры (x) и осадкам (z) достаточно точный.

Таблица 5.6

Расчет данных для критерия хи-квадрат

y	y'	$y - y'$	$(y - y')^2$	$\frac{(y - y')^2}{y'}$
300	302,2	–2,26	5,11	0,017
350	341,7	8,31	69,06	0,202
370	373,1	–3,09	9,55	0,026
420	423,7	–3,71	13,76	0,032
450	458,6	–8,62	74,30	0,162
500	491	8,97	80,46	0,164
		Σ – 0,4	252,241	$\chi^2 = 0{,}603$

Для определения ошибки уравнения линейной множественной регрессии показатели рассчитываем по формулам (5.12–5.14):

$$\Sigma a_y^2 = 978\ 300 – 6 \cdot 398{,}33^2 = 29\ 299{,}4;$$

$$\Sigma a_y a_x = 40\ 571 – 6 \cdot 398{,}33 \cdot 16{,}68 = 706{,}14;$$

$$\Sigma a_y a_z = 275\ 600 – 6 \cdot 398{,}33 \cdot 112 = 7922{,}3.$$

Затем подставляем полученные значения в формулу (5.11):

$$m = \sqrt{\frac{26\ 299{,}4 – (5{,}01 \cdot 706{,}14 + 2{,}84 \cdot 7922{,}3)}{6 – 3}} = 9{,}35 \text{ г/м}^2.$$

Таким образом, прогнозируя урожай биомассы трав за период вегетации по температуре и осадкам, мы рискуем ошибиться в среднем на 9,35 г/м^2, т. е. на 2,3 %.

Уравнения регрессии широко используются в некоторых научных исследованиях и в практических целях.

Глава 6. ФАКТОРНЫЙ АНАЛИЗ

6.1. СУЩНОСТЬ И ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ

При изучении влияния факторов, включающих много признаков, на исследуемый объект используют метод многомерного статистического анализа, в частности, факторного анализа. Например, влияние животноводческих комплексов на загрязнение окружающих грунтовых вод.

Факторный анализ основывается на использовании статистических знаний (вычислении стандартных отклонений, знании корреляционного и регрессионного анализов). В большинстве случаев исследуется система корреляций, отраженных в корреляционной матрице. Факторный анализ представляет собой ветвь математической статистики, цель которого – разработка моделей, понятий и методов, позволяющих анализировать и интерпретировать массивы экспериментальных данных независимо от их физической природы. Анализ данных включает краткое описание распределения объектов, установление взаимоотношения процессов и явлений, отражающихся в виде *параметров.*

Используемый набор моделей и методов предназначен для «сжатия» информации, содержащейся в корреляционной матрице. В основе различных моделей факторного анализа лежит следующая гипотеза: признаки являются косвенными характеристиками объекта или явления, представляя в совокупности тот или иной фактор. В связи с этим задача факторного анализа состоит в том, чтобы показать наблюдаемые признаки в виде линейных комбинаций факторов. Признаки, входящие в одну и ту же группу, сильно коррелируют между собой. Задача выявления факторов понимается как разбиение признаков на группы таким образом, чтобы можно было описать взаимоотношения между ними.

Разработано несколько вариантов факторного анализа с использованием коэффициентов только линейной корреляции (нелинейная корреляция вызывает затруднения при обработке материала). Среди них наиболее употребительны *метод главных компонент, метод главных факторов* и *центроидный метод.* Определение главных компонент и главных факторов производится с помощью ЭВМ.

Наиболее типичной формой представления данных является *матрица.* Это прямоугольная (или квадратная) таблица чисел, вертикальный ряд которой (столбец) обозначается индексом *j*, горизонтальный (строка) – индексом *i.* Любой элемент матрицы обозначается символом *a* с индексами, первый указывает номер строки, второй – номер столбца, которым соответствует данный элемент (в общем виде a_{ij}). Матрица обозна-

чается прописной буквой (*A*, *B* и т. д.). О матрице, имеющей *m* строк и *n* столбцов, говорят, что ее порядок составляет *m·n.* Квадратная матрица *n·n* имеет порядок *n.* В общем виде матрица записывается следующим образом:

$$A = \begin{vmatrix} a_{11}a_{12}a_{13}\ldots a_{1n} \\ a_{21}a_{22}a_{23}\ldots a_{2n} \\ a_{m1}a_{m2}a_{m3}\ldots a_{mn} \end{vmatrix}.$$

В факторном анализе с использованием правил матричной алгебры часто встречается операция умножения матриц. Для того чтобы умножить матрицу *A* на матрицу *B*, необходимо следующее условие: матрица *A* должна иметь столько столбцов, сколько строк в матрице *B*. Сам процесс умножения протекает по правилу «строка на столбец». Это правило означает, что каждый элемент матрицы произведения представляет собой сумму произведений элементов строки первой матрицы на соответствующие элементы столбца второй матрицы, например:

$$A = \begin{vmatrix} a_{11}a_{12} \\ a_{21}a_{22} \\ a_{31}a_{32} \end{vmatrix} \times B = \begin{vmatrix} b_{11}b_{12}b_{13} \\ b_{21}b_{22}b_{23} \end{vmatrix} = C =$$

$$= \begin{vmatrix} (a_{11}b_{11} + a_{12}b_{21})(a_{11}b_{12} + a_{12}b_{22})(a_{11}b_{13} + a_{12}b_{23}) \\ (a_{21}b_{11} + a_{22}b_{21})(a_{21}b_{12} + a_{22}b_{22})(a_{21}b_{13} + a_{22}b_{23}) \\ (a_{31}b_{11} + a_{32}b_{21})(a_{31}b_{12} + a_{32}b_{22})(a_{31}b_{13} + a_{32}b_{23}) \end{vmatrix}.$$

Матрица-произведение будет иметь всегда столько строк, сколько их было в первой матрице, и столько столбцов, сколько их было во второй матрице: $(p \cdot q) \cdot (q \cdot r) = (p \cdot r)$.

Существует ряд математических методов, которые по информации, заложенной в матрице, позволяют провести классификацию объектов. Такие методы объединены в многомерный анализ, а при наличии одной строки в матрице – в одномерный анализ.

В факторном анализе используются следующие виды матриц: *диагональная* (в ней отличны от нуля только элементы, лежащие на главной диагонали), *скалярная* (все элементы диагональной матрицы равны между собой), *единичная* (все элементы главной диагонали равны единице), *обратная* (аналогична обратному числу в арифметике).

Элементами исходной матрицы в факторном анализе являются коэффициенты корреляции. В ходе анализа вычисляется также общая дисперсия σ^2, указывающая, в каких границах находятся значения признаков,

которые характеризуют фактор. Кроме общей дисперсии в анализе учитывается факторная дисперсия (общность) и специфическая дисперсия, связанная с некоторой переменной и характеризующая только ее. Дисперсию, обусловленную ошибкой, стремятся свести к минимуму.

В итоге составляется факторная матрица. Элементы столбцов матрицы представляют собой *факторные нагрузки,* или *коэффициенты факторного отображения,* выраженные коэффициентами корреляции данной переменной с данным фактором. Таким образом, коэффициенты факторного отображения характеризуют фактор и его влияние на объект.

Результат факторного анализа можно также выразить в виде графика, который наглядно иллюстрирует полученные выводы. Каждую из двух связанных друг с другом переменных можно изобразить как вектор, т. е. отрезок прямой, имеющий определенную длину и направление. Величина корреляции между переменными равна произведению абсолютных величин обоих векторов на косинус угла между ними: $r_{1,2} = h_1 h_2 \cos \alpha_{1,2}$, где $r_{1,2}$ – коэффициент корреляции; h_1 – длина вектора, соответствующая первой переменной 1; h_2 – длина вектора, соответствующая переменной 2; $\cos \alpha_{1,2}$ – угол между векторами h_1 и h_2.

6.2. ПОСЛЕДОВАТЕЛЬНОСТЬ ОПЕРАЦИЙ

На конкретном примере рассмотрим один из методов факторного анализа. На основе выборки по 395 ландшафтам в пределах водораздельного пространства была получена исходная информация о восьми признаках агроландшафта. Они включают: 1) органические удобрения; 2) минеральные удобрения; 3) известь; 4) пестициды; 5) содержание гумуса в пахотном горизонте; 6) реакцию среды; 7) влажность почвы; 8) содержание физической глины. Следует определить, какова роль этих признаков в эволюции агроландшафтов.

Первый этап. Производится вычисление коэффициентов корреляции между всеми изучаемыми параметрами (табл. 6.1). Корреляционная матрица R симметрична, поэтому достаточно заполнить лишь ее половину до линии диагонали. Если показатель коррелирует сам с собой, коэффициент корреляции равен единице.

Второй этап. Для описания параметров используется линейная модель (параметры выражаются через скрытые гипотетические факторы линейно).

Таблица 6.1

Корреляционная матрица *R* для признаков агроландшафта

Показатели	1	2	3	4	5	6	7	8
1. Органические удобрения	1							
2. Минеральные удобрения	0,846	1						
3. Известь	0,805	0,881	1					
4. Пестициды	0,859	0,826	0,801	1				
5. Гумус	0,473	0,376	0,380	0,436	1			
6. Реакция почвы	0,398	0,326	0,319	0,329	0,762	1		
7. Влажность почвы	0,301	0,277	0,237	0,327	0,730	0,583	1	
8. Физическая глина	0,382	0,415	0,345	0,365	0,629	0,577	0,539	1

Примечание. В столбцах приведены признаки, аналогичные указанным в строках.

Основная модель факторного анализа может быть записана в виде формулы:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jm}F_m + d_j u_{ji},$$

где z_j – признак, F_1 – фактор, a_{ji} – приближение (коэффициент) факторного отображения (нагрузки). Первый член правой части равенства показывает долю первого фактора в исследуемых явлениях, второй – долю второго фактора, последний – долю независимого фактора (остаток). Чем больше величина коэффициента факторного отображения при факторе, тем больше роль данного фактора в рассматриваемом явлении.

Для выражения общей дисперсии определяется факторная дисперсия, или *значение общности* (σ_i^2) для каждого диагонального признака. Наиболее простой способ ее установления – вычисление первого центроидного фактора (табл. 6.2). На главную диагональ корреляционной матрицы помещают максимальные значения коэффициентов корреляции каждого столбца. Отношение квадрата суммы элементов каждого столбца к сумме всех элементов матрицы составит факторную дисперсию для столбца j:

$$\sigma_j^2 = \frac{\left(\sum_{i=1}^{n} r_j\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij}}, \qquad (6.1)$$

где Σr_i – суммарный коэффициент корреляции по столбцу; $\Sigma\Sigma r_{ij}$ – сумма восьми суммарных коэффициентов корреляции.

Таблица 6.2

Корреляционная матрица *R* с приближенными значениями общностей

Номер признака	1	2	3	4	5	6	7	8
1	**0,859**	0,846	0,805	0,859	0,473	0,398	0,301	0,382
2	0,846	**0,881**	0,881	0,826	0,376	0,326	0,277	0,415
3	0,805	0,881	**0,881**	0,801	0,380	0,319	0,237	0,345
4	0,859	0,826	0,801	**0,859**	0,436	0,329	0,327	0,365
5	0,473	0,376	0,380	0,436	**0,762**	0,762	0,730	0,629
6	0,398	0,326	0,319	0,329	0,762	**0,762**	0,583	0,577
7	0,301	0,277	0,237	0,327	0,730	0,583	**0,730**	0,539
8	0,382	0,415	0,345	0,365	0,629	0,577	0,539	**0,629**
Σr_j	4,923	4,828	4,649	4,802	4,548	4,056	3,724	3,881
								$\Sigma\Sigma r_{ij}$ = 35,411

Подставив данные в формулу (6.1), имеем первую факторную дисперсию: $\sigma_i^2 = 4{,}923^2 / 35{,}411 = 0{,}684$. Аналогично проводим расчет дисперсии по остальным столбцам табл. 6.2.

Полученные данные помещаем по главной диагонали редуцированной корреляционной матрицы R^x (табл. 6.3).

Если рассчитанные коэффициенты корреляции мало отличаются от исходных, значит, модель хорошо описывает экспериментальные данные. Однако максимальный коэффициент $r_1 = 0{,}859$ (см. табл. 6.2) отличается от рассчитанного $r_1 = 0{,}684$ (см. табл. 6.3).

Третий этап. Проводим группировку параметров с целью определения факторов. Восемь параметров образуют две группы (см. табл. 6.1): первые четыре параметра характеризуют химическую мелиорацию почв (первый фактор), остальные – их плодородие (второй фактор).

Четвертый этап. Находим первое приближение факторного отображения. Предполагается, что полученные факторы не коррелируют между собой. Для каждой строки матрицы R^x вычисляем сумму коэффициентов корреляции:

$$\Sigma r_{i1} = r_{i1} + r_{i2} + \ldots r_{ij}, \qquad (6.2)$$

где $\Sigma r_1 = 0{,}854 + 0{,}846 + 0{,}805 + 0{,}859 + 0{,}473 + 0{,}398 + 0{,}301 + 0{,}382 = 4{,}738$ (см. табл. 6.3).

Результаты записываем в предпоследний столбец редуцированной корреляционной матрицы. Каждую сумму Σr_i делим на максимальное значение (в нашем примере максимальная $\Sigma r_1 = 4{,}388$). Имеем первое приближение $a_{ij}^{(1)}$ факторного отображения:

$$a_{11}^{(1)} = 4{,}738 : 4{,}738 = 1{,}000;\ a_{21}^{(1)} = 4{,}605 : 4{,}738 = 0{,}971.$$

Таблица 6.3

Редуцированная корреляционная матрица R^x

Номер признака	1	2	3	4	5	6	7	8	$\Sigma r_i^{(1)}$	$a_{ij}^{(1)}$
1	**0,684**	0,846	0,805	0,849	0,473	0,398	0,301	0,382	4,738	1,000
2	0,846	**0,658**	0,881	0,826	0,376	0,326	0,277	0,415	4,605	0,971
3	0,805	0,881	**0,610**	0,801	0,380	0,319	0,237	0,345	4,378	0,924
4	0,859	0,826	0,801	**0,651**	0,436	0,329	0,327	0,365	4,595	0,969
5	0,473	0,376	0,380	0,436	**0,584**	0,762	0,730	0,629	4,370	0,922
6	0,398	0,326	0,319	0,329	0,762	**0,464**	0,583	0,577	3,758	0,793
7	0,301	0,277	0,237	0,327	0,730	0,583	**0,391**	0,539	3,391	0,715
8	0,382	0,415	0,345	0,365	0,629	0,577	0,539	**0,425**	3,677	0,776

Данные вносим в последний столбец редуцированной матрицы.

Пятый этап. Возводим редуцированную матрицу (см. табл. 6.3) в квадрат. Для этого необходимо каждое число возвести в квадрат в первом столбце матрицы и суммировать результаты:

$$(0,684)^2 + (0,846)^2 + (0,805)^2 + (0,859)^2 + (0,473)^2 +$$
$$+ (0,398)^2 + (0,301)^2 + (0,382)^2 = 3,188.$$

Получаем первый элемент матрицы R^2 (табл. 6.4). Поскольку квадрат симметричной матрицы есть также симметричная матрица, то вычисляем диагональные элементы и элементы выше (или ниже) диагонали. Для контроля выполненных вычислений суммируем элементы строк $\Sigma r_i^{(2)}$ матрицы R^2, например:

$$3,188 + 3,450 + 3,301 + 3,325 + 2,577 + 2,185 + 1,903 + 2,179 = 22,11.$$

Таблица 6.4

Квадрат корреляционной матрицы

Номер признака	1	2	3	4	5	6	7	8	$\Sigma r_i^{(2)}$	$T_i^{(2)}$	$a_{ij}^{(2)}$
1	3,450	3,450	3,301	3,325	2,577	2,185	1,903	2,179	22,11	22,26	1
2	3,450	3,475	2,322	3,333	2,471	2,093	1,815	2,115	22,07	22,07	0,986
3	3,301	3,322	3,181	3,188	2,341	1,983	1,718	2,003	21,03	21,03	0,94
4	3,325	3,333	3,188	3,213	2,461	2,084	1,810	2,085	21,49	21,49	0,961
5	2,577	2,471	2,341	2,461	2,966	2,550	2,780	2,372	20,01	20,01	0,894
6	2,185	2,093	1,983	2,084	2,550	2,200	1,965	2,041	17,10	17,10	0,764
7	1,903	1,815	1,718	1,810	2,278	1,965	1,765	1,820	15,07	15,07	0,673
8	2,176	2,115	2,003	2,085	2,372	2,041	1,820	1,925	16,54	16,54	0,739

Затем определяем сумму элементов строк с помощью формулы:

$$T_i^{(2)} = \sum_{i||1}^{n} r_i^{(2)} r_{i/j}, \qquad (6.3)$$

где $\Sigma r_i^{(2)}$ – сумма элементов строк матрицы;

$$r_{i/j} = \sum_{i=1}^{n} r_i / \sum_{j=1}^{m} r_j.$$

Приведем пример расчета указанных выше показателей: $r_{1/1}$ = 4,738 : : 4,923 = 0,962; $T_1^{(2)}$ = 22,11 · 0,962 = 22,26 (см. табл. 6.4). Значение $T_i^{(2)}$ должно соответствовать величине $r_i^{(2)}$.

Величина $a_{ij}^{(2)}$ представляет собой второе приближение чисел, которое получаем путем деления каждой из сумм $T_i^{(2)}$ на максимальную величину в данном столбце (см. табл. 6.4). Показатель $a_{ij}^{(2)}$ сравниваем с соответствующими значениями первого приближения $a_{ij}^{(1)}$ (см. табл. 6.3) (различие между ними должно быть < 0,005). Различие $a_{i3}^{(1)}$ и $a_{i3}^{(2)}$ результатов превышает 0,005, поэтому возведение корреляционной матрицы в квадрат производится до тех пор, пока собственный вектор не перестанет изменяться. Перед очередной операцией возведения матрицы R в степень вычисляются значения $T_i^{(e)}$ и $a_{ij}^{(e)}$ последующей матрицы R^e. Если коэффициенты a_{ij} последующей матрицы совпадают с коэффициентами предыдущей матрицы с достаточной точностью, то нет необходимости вычислять остальные элементы матрицы R^e. В нашем случае максимальное различие между $a_{7j}^{(8)}$ и $a_{7j}^{(4)}$ составляет всего 0,004, поэтому элементы R^e можно не вычислять (табл. 6.5).

Таблица 6.5

Показатели четвертой и восьмой степени корреляционной матрицы

Номер признака	$T_i^{(4)}$	$a_{ij}^{(4)}$	$T_i^{(8)}$	$a_{ij}^{(8)}$
1	447,9	1,000	176,7	1,000
2	443,0	0,989	174,8	0,989
3	422,4	0,943	166,7	0,943
4	430,7	0,961	19,9	0,961
5	390,9	0,872	153,7	0,869
6	333,6	0,744	131,2	0,742
7	293,6	0,655	115,4	0,651
8	324,0	0,723	127,5	0,721

Шестой этап. Вычисляем коэффициенты при первом факторе F_1. Найденное восьмое приближение чисел $a_{7j}^{(8)}$ (см. табл. 6.5) представляет собой вектор и умножается на R. Значение R_{q1}, соответствующее $a_{11}^{(8)} = 1{,}000$, является первым корнем характеристического уравнения λ_k. Далее рассчитываем коэффициенты b_{i1} при первом факторе F_1 (табл. 6.6), которые учитывают максимально возможную долю суммарной общности:

$$b_{i1} = a_{i1}\sqrt{\lambda_1} / \sqrt{\Sigma a_{i1}^{(2)}}, \tag{6.4}$$

где $a_{i1} = a_{11}^{(8)}$; $\Sigma a_{i1}^2 = \Sigma a_{i1}^{2(2)}$; $\lambda_1 = \sum_{i=1}^{n} b_{i1}^2$; $b_{i1} = 0{,}85827$.

Получим искомые коэффициенты b_{i1} при F_1 в факторном отображении. Сумма вкладов первого фактора в суммарную общность должна быть равна первому характеристическому корню:

$$\sum_{i=1}^{n} b_{i1}^2 = \lambda_1. \tag{6.5}$$

В нашем примере $\lambda_1 = 4{,}4556$, $\sum_{i=1}^{8} b_{i1}^2 = 4{,}455$; следовательно, результаты являются удовлетворительными.

Таблица 6.6

Квадрат корреляционной матрицы

Номер признака	$a_{i1}^{(8)}$	R_{q1}	b_{i1}
1	1,000	4,455	0,858
2	0,989	4,408	0,849
3	0,943	4,208	0,810
4	0,961	4,285	0,825
5	0,869	3,875	0,747
6	0,742	3,307	0,637
7	0,653	2,909	0,561
8	0,721	3,214	0,619
			$D_1 = 5{,}905$

Седьмой этап. Проводим поиск фактора, который учитывал бы максимум остаточной общности. Для этого после учета F_1 необходимо построить матрицу R_1, используя коэффициенты первого фактора. По строкам табл. 6.7 рассчитываются суммы элементов E_{i1}. Например,

$E_{11} = 0{,}736 + 0{,}728 + 0{,}695 + 0{,}708 + 0{,}641 + 0{,}547 + 0{,}481 + 0{,}531 = 5{,}067$.

Результаты сравниваем с произведениями $b_{i1}D_1$, где $D_1 = \Sigma b_{i1} = 5{,}905$ (см. табл. 6.6).

Таблица 6.7

Матрица произведений $\tilde{R}_1(a_{i1}a_{j1})$

Номер признака	1	2	3	4	5	6	7	8	E_{i1}	$b_{i1}D_1$
1	0,736	0,728	0,695	0,708	0,641	0,547	0,481	0,531	5,067	5,067
2	0,728	0,721	0,688	0,700	0,634	0,541	0,476	0,526	5,014	5,014
3	0,695	0,688	0,656	0,668	0,605	0,516	0,454	0,501	4,783	4,784
4	0,708	0,700	0,668	0,681	0,616	0,526	0,463	0,511	4,873	4,782
5	0,641	0,634	0,605	0,616	0,558	0,476	0,419	0,462	4,411	4,412
6	0,547	0,541	0,516	0,526	0,476	0,406	0,357	0,394	3,763	3,762
7	0,481	0,476	0,454	0,463	0,419	0,357	0,315	0,347	3,312	3,313
8	0,531	0,526	0,501	0,511	0,462	0,394	0,347	0,383	3,655	3,656

Первые остаточные коэффициенты корреляции (табл. 6.8) равны разности соответствующих элементов матриц R^x и R_1 (см. табл. 6.3 и 6.7).

Суммы элементов матрицы R_1, полученной по строкам, должны быть равны разности соответствующих сумм матриц R^x и R_1. После выполнения операций по первому фактору (табл. 6.9) переходим к второму фактору, сведения приведены в табл. 6.10. В итоге получаем коэффициенты факторного отображения и общности (табл. 6.11), по которым делаем соответствующие выводы.

Таблица 6.8

Матрица первых остаточных коэффициентов корреляции R_1

Номер признака	1	2	3	4	5	6	7	8	Σr_{i1}	$a_{i1}^{(1)}$
1	0,118	0,118	0,110	0,151	–0,168	–0,149	–0,180	–0,149	–0,149	–0,608
2	0,118	0,176	0,193	0,126	–0,258	–0,215	–0,199	–0,111	–0,170	–0,693
3	0,110	0,193	0,177	0,133	–0,225	–0,197	–0,217	–0,156	–0,182	–0,742
4	0,151	0,126	0,133	0,102	–0,180	–0,197	–0,136	–0,146	–0,147	–0,600
5	–0,168	–0,258	–0,225	–0,180	0,312	0,286	0,311	0,167	0,245	1,000
6	–0,149	–0,215	–0,197	–0,197	0,286	0,281	0,226	0,183	0,218	0,889
7	–0,180	–0,199	–0,217	–0,136	0,311	0,226	0,206	0,192	0,203	0,828
8	–0,149	–0,111	–0,156	–0,146	0,167	0,183	0,192	0,196	0,176	0,718

Таблица 6.9

Этапы вычисления приближенных значений коэффициентов

Номер признака	Σr_{ij}	E_{i1}	Σr_{i1}	$a_{i1}^{(1)}$	$\lambda_1 \Sigma r_1$	$\Sigma r_1^{(2)}$	$\Sigma r_{i1}^{(2)}$	$a_{i1}^{(2)}$
1	4,918	5,067	–0,149	–0,608	22,37	22,57	–0,20	–0,57
2	4,844	5,014	–0,170	–0,693	22,07	22,33	–0,26	–0,73
3	4,601	4,783	–0,182	–0,742	21,04	21,30	–0,26	–0,73
4	4,726	4,873	–0,147	–0,600	21,50	21,70	–0,20	–0,57
5	4,656	4,411	0,245	1,000	20,02	19,65	0,37	1,00
6	3,981	3,763	0,218	0,889	17,10	16,76	0,34	0,89
7	3,521	3,312	0,203	0,828	15,07	14,75	0,32	0,84
8	3,831	3,655	0,176	0,718	16,54	16,28	0,26	0,68

Таблица 6.10

Вычисление коэффициентов при факторе F_2

Номер признака	a_{i2}	R_{1q2}	a_{i2}	R_{1q2}	a_{i2}	R_{1q2}	a^2_{i2}	b_{i2}
1	–0,57	–0,865	–0,580	–0,8856	–0,5851	–0,8852	–0,5851	–0,328
2	–0,73	–1,100	–0,737	–1,1152	–0,7368	–1,1148	–0,7369	–0,414
3	–0,73	–1,097	–0,735	–1,1118	–0,7345	–1,1112	–0,6345	–0,412
4	–0,57	–0,902	–0,605	–0,9129	–0,6031	–0,9129	–0,6034	–0,339
5	1,00	1,492	1,000	1,5136	1,0000	1,5129	1,0000	0,561
6	0,89	1,348	0,903	1,3666	0,9029	0,3660	1,9029	0,507
7	0,84	1,300	0,871	1,3144	0,8684	0,3142	1,8687	0,488
8	0,68	0,987	0,662	1,0050	0,6610	0,0001	1,6610	0,371

Таблица 6.11

Этапы вычисления приближенных значений коэффициентов

Признаки	Коэффициенты факторного отображения			Общность	
	b_{i1}	b_{i2}	u_i	исходная	вычисленная
1. Органические удобрения	0,858	–0,328	0,395	0,854	0,844
2. Минеральные удобрения	0,849	–0,414	0,328	0,897	0,892
3. Известь	0,810	–0,412	0,417	0,833	0,826
4. Пестициды	0,825	–0,339	0,452	0,783	0,796
5. Гумус	0,747	0,567	0,375	0,870	0,873
6. Реакция почвы	0,637	0,507	0,581	0,687	0,663
7. Влажность почвы	0,561	0,488	0,669	0,521	0,553
8. Физическая глина	0,619	0,371	0,692	0,579	0,521
Сумма				6,024	5,968
Вклад факторов	4,455	1,511			
Процент от суммарной исходной общности	74,00	25,10			99,10

Поскольку коэффициенты при первом факторе (b_{i1}) положительные и достаточно велики, можно утверждать, что роль первого фактора (химическая мелиорация) в эволюции агроландшафтов весьма существенна. Второй фактор (b_{i2} – плодородие почв) относится к биполярным, так как имеет одинаковое число положительных и отрицательных нагрузок: коэффициенты со знаком плюс соответствуют признакам, отражающим степень плодородия почв, со знаком минус – признакам, отражающим химическую мелиорацию. Таким образом, эволюция агроландшафтов обусловлена прежде всего химической мелиорацией почв. Признаки плодородия почв формируются под воздействием первого фактора комплексной химической мелиорации и в эволюции агроландшафтов выполняют второстепенную роль.

Из всех признаков наибольший удельный вес в эволюции агроландшафтов занимают органические удобрения (b_{i1} = 0,858). Коэффициенты факторного отображения второго фактора, выраженные отрицательными числами, характерны для показателей, описывающих степень химизации почв. Это позволяет интерпретировать полученные данные как дефицит химических мелиорантов для рассматриваемых конкретных условий, что отражается отрицательно на прогрессивной эволюции агроландшафтов.

Изложенные выводы подтверждаются данными рис. 6.1. Судя по размещению коэффициентов факторного отображения, признаки 1–4 (хими-

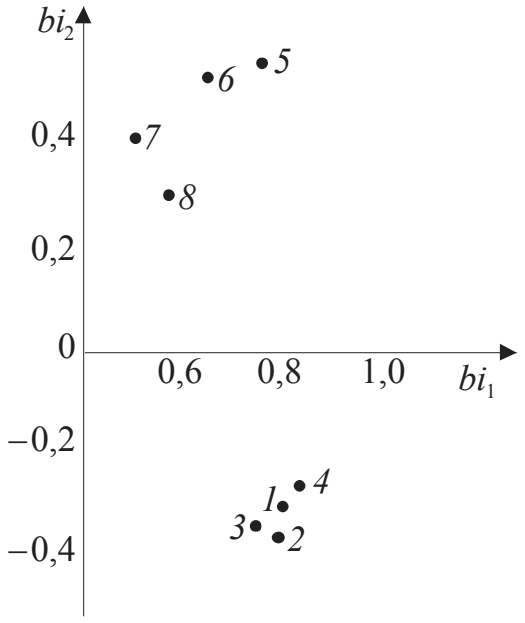


Рис. 6.1. Распределение коэффициентов факторного отображения

ческие мелиоранты) расположены компактно в пространстве, что указывает на их важную совместную роль в эволюции агроландшафтов. Между признаками 5–8, отражающими степень плодородия почв, связь слабее и соответственно слабее их влияние на эволюцию агроландшафтов.

Метод факторного отображения используется при решении многих землеустроительных задач: районирования, количественной оценки влияния природных условий на сельскохозяйственное производство и др. По значениям факторов можно составить картосхему, отображающую территориальное выражение исследуемых факторов.

Глава 7. МЕТОДЫ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

Важнейшей задачей агропромышленного комплекса является повышение эффективности эксплуатации земельных угодий. Решающее значение в этом принадлежит землеустройству. Необходимо находить оптимальные варианты трансформации земель и устройства территории, решать задачи, связанные с планированием использования земельных, материальных, трудовых и денежных ресурсов. Переработка большого объема информации и научно-обоснованные решения по этим вопросам невозможны без применения экономико-математических методов моделирования на базе интенсивно разрабатываемых в последние годы методов линейного программирования.

Основными задачами методов линейного программирования являются:

– обработка информации с помощью экономико-статистических методов для нужд землеустройства;

– изучение оптимизационных методов линейного программирования, позволяющих решать землеустроительные задачи в условиях ограниченных ресурсов и отыскать резервы для повышения эффективности организации производства и территории.

Линейные модели активно используются также в экономике и экономической географии, как достаточно эффективные в ряде ситуаций. Линейная функция (тройное правило) самая удобная, простая, хорошо разработанная математическая модель.

Линейность – это свойство математических выражений и функций. Выражение типа $ax + by$, где x, y – переменные величины, a, b – постоянные числа, называется линейным относительно переменных x, y. Если переменных больше двух ($x_1, x_2, \ldots x_n$), линейное выражение относительно их имеет вид: $a_1x_1 + a_2x_2 + \ldots a_nx_n$. В линейное выражение все переменные входят в первой степени и никакие переменные не перемножаются.

Линейное программирование – это совокупность методов решения экстремальных задач, в которых цель (критерий оптимальности) и условия (ограничения) заданы уравнениями и неравенствами первой степени. Программирование используется в данной ситуации как планирование, линейное – означает, что ищется экстремум линейной целевой функции при линейных ограничениях (линейных уравнениях, линейных неравенствах). Однако вычислительные средства при решении задач этого класса играют существенную роль в повышении эффективности их приложений.

Для решения задач с применением линейного программирования эффективны следующие:

● составление смеси продукции предполагает выбор наиболее экономичного топлива, пищевых продуктов и т. д.;

● задачи производства – подбор наиболее выгодной производственной программы выпуска одного или нескольких видов продукции при использовании некоторого числа ограниченных источников сырья;

● задачи распределения, или транспортные задачи;

● комбинированные задачи – производство товаров в разных местах, задачи производства и распределения объединяют в единую задачу.

Разработан ряд алгоритмов, среди которых наиболее известны *симплексный* и *распределительный методы*. Наиболее эффективен метод *эллипсоидов (графический)*. Оба метода базируются на последовательном улучшении первоначального плана путем повторения вычислений (интераций). После каждой интерации значение целевой функции улучшается. Процесс повторяется до получения оптимального плана, а полученный план проверяется на оптимальность разработанными простыми критериями.

Симплекс-метод более универсален, так как позволяет решать задачи, условия которых выражены в различных единицах измерения. В задачах, решаемых распределительным методом (транспортные задачи), все переменные должны иметь одну и ту же единицу измерения. Транспортные задачи являются специальной разновидностью симплекс-метода.

Землеустроительные задачи, решаемые методами линейного программирования, должны удовлетворять следующим требованиям:

● их решение не должно быть однозначным;

● иметь определенную целевую функцию, для которой ведется поиск максимального или минимального значения;

● иметь условия ограничения, формирующие область допустимых решений задачи.

7.1. СОСТАВНЫЕ ЧАСТИ ОБЩЕЙ МОДЕЛИ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

Все модели линейного программирования состоят из стандартных составных частей: совокупность основных переменных, линейные ограничения (условия), целевая функция, определяющая критерий оптимальности задачи.

Совокупность основных переменных характеризует размеры землепользований, площади, объемы производства, затраты материальных, трудовых, финансовых ресурсов.

Система линейных ограничений (условий) определяет область допустимых значений основных переменных. Каждое отдельное условие отражает реальное ограничение (нормы внесения удобрений, выполнение контрольных цифр бизнес-плана и т. д.).

Целевая функция представляется показателем, который обобщенно характеризует один из аспектов деятельности хозяйства данной землеустроительной задачи, например, чистый доход, валовую продукцию и т. д.

Критерий оптимальности в зависимости от условий задачи требует максимизации или минимизации целевой функции при заданных ограничениях.

В общем виде формализованная модель линейного программирования, построенная для решения задачи, в которой выделено N основных переменных $x_1, ..., x_N$ и M ограничений, будет иметь следующий вид.

Целевая функция

$$F = (c_1x_1 + c_2x_2 + \ldots + c_Nx_N) \rightarrow \max\ (\min) \qquad (7.1)$$

Система ограничений:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1N}x_N \therefore b_1; \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2N}x_N \therefore b_2; \\ ..... \\ a_{M1}x_1 + a_{M2}x_2 + ... + a_{MN}x_N \therefore b_M, \end{cases}$$

где знак $\therefore$ означает или $\leq$, или $\geq$, или =; константы $b_1 \ldots b_M$ в правых частях ограничений предполагаются неотрицательными; требование неотрицательности основных переменных: $x_1 \geq 0,\ x_2 \geq 0,\ \ldots,\ x_N \geq 0$.

Совокупность ограничений и требований неотрицательности основных переменных определяет область допустимых значений задачи.

В кратком виде вышеприведенная развернутая запись будет следующей:

$$F = \sum_{j=1}^{N}(c_j x_j) \longrightarrow \max(\min); \sum_{j=1}^{N}(a_{ij}x_j) \therefore b_i;\ i = 1, ..., M;$$

$$x_j \geq 0,\ \ j = 1, ..., N. \qquad (7.2)$$

Распределительный метод среди задач линейного программирования получил распространение из-за упрощения расчетов, точности вычислений и снижения затрат времени на ввод исходной информации. Метод предложили А. Толстой и Л. В. Конторович в 1939–1940 гг. Первоначально он применялся в задачах, связанных с транспортировкой грузов, их распределением между поставщиком и потребителем, поэтому получил название «транспортная задача». Они могут применяться при решении ряда землеустроительных задач: распределение севооборотов и угодий по участкам, размещение культур на землях различных категорий, перераспределение участков между хозяйствами для экономии транспортных затрат и др.

Суть распределительной задачи следующая. Заданы m источников ресурса (производители продукции, базы с готовой продукцией) и n пунктов его потребления. Запасы ресурса в источниках составляют A_i, $i = 1, \ldots, m$, потребности – B_j, $j = 1, \ldots, n$. Стоимость транспортировки единицы ресурса от i-го источника к j-му потребителю C_{ij}. Количество ресурса, транспортируемого от i-го источника к j-му потребителю X_{ij}. Требуется определить такие значения X_{ij}, при которых общие транспортные расходы будут минимальны.

При сбалансированности, когда общий спрос на запас ресурса у поставщиков и общий спрос на него у потребителя равны, задачу называют *закрытой*:

$$\sum_{i=1}^{m} A_i = \sum_{j=1}^{n} B_j \,. \tag{7.3}$$

Если баланс не выдерживается, то транспортная задача является *открытой*:

$$\sum_{i=1}^{m} A_i < \sum_{j=1}^{n} B_j \text{ , или } \sum_{i=1}^{m} A_i > \sum_{j=1}^{n} B_j \,. \tag{7.4}$$

При наличии баланса модель транспортной задачи формулируется следующим образом.

Целевая функция:

$$F = \sum (C_{ij} X_{ij}) \to \min. \tag{7.5}$$

Условия. Ограничения по запасам:

$$\sum_{j=1}^{n} X_{ij} = A_i, \; i = 1, ..., m. \tag{7.6}$$

Ограничения по потребностям:

$$\sum_{j=1}^{m} X_{ij} = B_j, \, j = 1, ..., n. \tag{7.7}$$

Условие баланса:

$$\sum_{i=1}^{m} A_i = \sum_{j=1}^{n} B_j. \tag{7.8}$$

Условие неотрицательности:

$$X_{ij} \geq 0, \, i = 1, \ldots, m, \; j = 1, \ldots, n. \tag{7.9}$$

Особенности распределительных транспортных задач следующие:

● условия задачи описываются уравнениями (в симплекс-методе описываются и неравенствами);

● все переменные выражаются в одних и тех же единицах измерения;

● во всех уравнениях коэффициенты при переменных равны единице;

● каждая переменная встречается только в двух уравнениях системы ограничений: в одном по строке (по запасам) и в одном по столбцу (по потребностям).

Целевая функция F выражает суммарные расходы на транспортировку грузов. Ограничения по запасам и по потребностям означают, что сумма ресурса, забираемого из i-го источника, должна быть равна запасу ресурса в нем, как и сумма ресурса, доставляемого j-му потребителю, должна быть равна его потребности.

Величина C_{ij} может выражать транспортные расходы (минимизация) или прибыль от транспортных операций (максимизация) и другие показатели.

Пример землеустроительной задачи, решаемой транспортным методом.

При землеустроительном обследовании в хозяйстве выделено 5 участков с различным плодородием, которые пригодны для трансформирования. Площади участков 250, 100, 520, 310 и 130 га. По проекту на них намечается разместить кормовой севооборот площадью 600 га, полевой – 560, улучшенные сенокосы – 150 га. Необходимо распределить севообороты и угодья по участкам так, чтобы получить максимальный чистый доход.

Матрицу исходных данных строим как в табл. 7.1. На «транспортном» языке эта задача может быть описана следующим образом. «Ресурсы» в источниках (A_i) – площади севооборотов и улучшенных сенокосов; «потребности в ресурсах» (B_j) – площади участков; «прибыль от транспортных операций» (C_{ij}) – чистый доход с единицы площади в центре клеток матрицы; «транспортируемый ресурс» (X_{ij}) – часть площади i-го севооборота или угодья, размещаемого на j-ом участке, которую придется распределить по клеткам матрицы с максимальными значениями C_{ij}; максимальная целевая функция (F_{max}) – чистый доход хозяйства от рационального размещения и трансформации угодий; $\sum a_i = \sum b_j$. Чистый доход проставляется в правом верхнем углу каждой клетки (C_{ij}, руб./га). Дальнейшее решение задачи проводится с использованием метода потенциалов.

Таблица 7.1

Исходные данные для землеустроительной задачи

Угодья и севообороты	b_j / a_i	Чистый доход при размещении на участке, руб./га (C_{ij})				
		1 пастбище, 250 га	2 пашня, 100 га	3 пашня, 520 га	4 пашня, 310 га	5 сенокосы, 130 га
Кормовой	600 га	800	1100	800	600	440
Полевой	560 га	1000	1800	2000	2200	2000
Улучшенные сенокосы	150 га	550	440	380	300	700

Для решения транспортных задач разработан ряд методов: функционала, потенциала, дельта-метод, лямбда-задача. Используются модифицированные модели: транспортно-производственная, многоэтапная, многопродуктовая.

Вначале рассмотрим основные правила работы с матрицей, составление и перемещение по цепи и расчет необходимых параметров.

Расположение элемента (числа) в матрице строго фиксировано. Строку обозначают буквой *i,* столбец – *j*, элемент матрицы – a_{ij} (где *i* – номер строки, *j* – номер столбца). Запись $a_{12,7}$ показывает, что данный элемент расположен в 12-ой строке и 7-ом столбце матрицы. Цифры, указывающие строку и столбец до 10 не разделяют запятой (a_{23}).

Матрицу обозначают заглавной буквой (*A, B, C*):

$$A \begin{bmatrix} a_{11}a_{12} \ldots a_{1n} \\ a_{21}a_{22} \ldots a_{2n} \\ \ldots\ldots\ldots\ldots \\ a_{m1}a_{m2} \ldots a_{mn} \end{bmatrix}$$

В матрице число строк равно *m*, столбцов – *n*. В сокращенном виде матрицу записывают $A = (a_{ij})$.

Размер матрицы определяют путем произведения *m* на *n*. Запись $\sum_{i=1}^{m} a_{ij}$ означает, что в матрице из чисел *a* необходимо просуммировать все числа матрицы по столбцам.

Матрицу можно транспонировать, т. е. перемещать элементы матрицы так, что ее строки становятся столбцами, а столбцы – строками. При большинстве вычислений (кроме умножения матрицы на матрицу) не имеет значения, что считать в ней строками, а что столбцами.

Вектор в матрице представляет собой упорядоченную последовательность элементов или ряд, состоящий из некоторого количества элементов. Поэтому вектором можно считать любую строку или любой столбец матрицы. Если размер матрицы $m \cdot n$, то она состоит либо из m векторов, в каждом из которых по n элементов, либо из n векторов, в каждом из которых по m элементов.

При решении транспортной задачи используются следующие обозначения:

i – индекс поставщика ($i = 1, 2, \ldots, m$);

j – индекс потребителя ($j = 1, 2, \ldots, n$);

a_i – мощность i-го поставщика;

b_j – спрос j-го потребителя;

C_{ij} – затраты на перевозку продукции от i-го поставщика j-му потребителю;

X_{ij} – количество продукции, которое необходимо перевезти от i-го поставщика j-му потребителю.

Условия транспортной модели приведены выше в составных частях общей модели линейного программирования. Совокупные затраты на перевозку сводятся к минимуму целевой функции. Исходная информация для решения транспортной задачи представлена в матрице:

Потребители		B_1	B_2	…	B_n
Поставщики	b_j a_i	b_1	b_2	…	b_n
A_1	a_1	C_{11} X_{11}	C_{12} X_{12}	…	C_{1n} $X_{1\,n}$
A_2	a_2	C_{21} X_{21}	C_{22} X_{22}	…	C_{2n} $X_{2\,n}$
…	…	…	…	…	…
A_m	a_m	C_{m1} X_{m1}	C_{m2} X_{m2}	…	C_{mn} X_{mn}

В транспортной задаче $m \cdot n > (m + n - 1)$ можно составить множество планов перевозок. Такие планы называют *допустимыми.*

В табл. 7.2 дана запись исходных данных задачи по 3-м поставщикам и 4-м потребителям. Указаны мощности поставщиков (a_i) и спросы потребителей (b_j), в правом верхнем углу клетки – затраты на перевозку единицы груза (C_{ij}).

По исходным данным табл. 7.2 могут быть составлены следующие допустимые планы перевозок (табл. 7.3).

Таблица 7.2

Исходные данные транспортной задачи

Поставщики, их мощности (a_i)		Потребители и их спрос (b_j)			
		B_1 **40**	B_2 **25**	B_3 **15**	B_4 **20**
A_1	**30**	5	4	1	2
A_2	**50**	1	2	3	4
A_3	**20**	3	2	5	1

Таблица 7.3

Допустимые планы перевозок грузов

a)

a_i \ b_j	**40**	**25**	**15**	**20**
30			25 (4)	5 (1)
50	40 (1)		10 (3)	
20				20 (1)

b)

a_i \ b_j	**40**	**25**	**15**	**20**
30			15 (1)	15 (2)
50	40 (1)	5 (2)		5 (4)
20		20 (2)		

c)

a_i \ b_j	**40**	**25**	**15**	**20**
30			15 (1)	15 (2)
50	25 (1)	20 (2)		5 (4)
20	15 (3)	5 (2)		

В допустимом плане C_{ij} обводятся кружком в случаях наличия в таких клетках поставок (X_{ij}), поэтому клетки таблиц с поставками условимся называть *клетками с кружками.*

В табл. 7.3 *а* число кружков 5, т. е. меньше чем $m + n - 1$; в табл. 7.3 *b* число кружков 6 (равно $m + n - 1$); в табл. 7.3 *с* кружков 7 (больше чем $m + n - 1$). В методе потенциалов число кружков в допустимом плане должно быть равно $m + n - 1$ с расположением их по *вычеркиваемой комбинации.*

Вычеркиваемая комбинация получается в случае, если каждый кружок – единственный в своем столбце или строке, и тогда он может быть вычеркнут. Такому условию соответствует распределение в табл. 7.3 *b*. Последовательность вычеркивания следующая: 40; 15; 20; 15; 5; 5 или 20; 40; 15; 5; 15; 5.

Существует несколько способов составления допустимого (базисного) плана: северо-западного угла, поисков наименьшего элемента в столбце, наименьшего элемента в строке, наименьшего элемента в матрице. Роль наименьшего элемента выполняет C_{ij} (цифры в правом верхнем углу клеток матрицы).

Способ северо-западного угла более сложный, и в случае большой матрицы не рекомендуется его использование. Если столбцов меньше, используют поиски наименьшего C_{ij} в столбцах; если строк мало – способ поиска наименьшего элемента в строках; если матрица большая, проводится поиск наименьшего элемента в клетках матрицы.

Проведем распределение поставок перечисленными выше способами при одинаковых исходных данных и для сравнения вычислим их функционалы: $F = \sum_{i=1}^{m}\sum_{j=1}^{n}(C_{ij}X_{ij}) \to \min$.

Способ северо-западного угла. В матрице $m + n - 1 = 7$. Поставки распределяем по диагонали не зависимо от величины C_{ij}: $30^5 - 25^2 - 15^5 - 10^3$. Остаток поставок распределяем между потребителями с учетом минимальных значений C_{ij}: $- 5^3 - 5^4$.

b_i / a_i	B_1 40	B_2 25	B_3 15	B_4 10
A_1 30	⑤ 30	2	3	4
A_2 30	④ 5	② 25	2	① 0
A_3 20	③ 5	3	⑤ 15	2
A_4 10	2	4	6	③ 10

Все потребители получили необходимый объем продукции.

Функционал при таком способе распределения имеет величину:

$$F = (30 \cdot 5) + (25 \cdot 2) + (15 \cdot 5) + (10 \cdot 3) + (5 \cdot 4) + (5 \cdot 3) = 320.$$

Способ поиска наименьшего C_{ij} в столбце. Поставки распределяются последовательно по столбцам с учетом наименьших C_{ij}. Получаем следующую последовательность:

$$10^2 – 20^3 – 10^4 – 25^2 – 15^2 – 5^1 – 5^4.$$

Расчет функционала:

$$F = (10 \cdot 4) + (20 \cdot 3) + (10 \cdot 2) + (25 \cdot 2) + (15 \cdot 2) + (5 \cdot 4) + (5 \cdot 1) = 225.$$

Полученный функционал (F = 225) указывает, что допустимый план по способу наименьшего элемента в столбе более оптимальный, чем по способу северо-западного угла (F = 320).

b_j / a_i	B_1 40	B_2 25	B_3 15	B_4 10
A_1 30	5	② 25	3	④ 5
A_2 30	④ 10	2	② 15	① 5
A_3 20	③ 20	3	5	2
A_4 10	② 10	4	6	3

Способ поиска наименьшего элемента C_{ij} в строке. Поставки распределяем последовательно сверху вниз по строкам с учетом наименьших величин C_{ij}: $25^2 – 10^1 – 20^3 – 10^2 – 15^2 – 5^5$. Получаем функционал 215.

b_j / a_i	B_1 40	B_2 25	B_3 15	B_4 10
A_1 30	⑤ 5	② 25	3	4
A_2 30	④ 5	2	② 15	① 10
A_3 20	③ 20	3	5	2
A_4 10	② 10	4	6	3

По данному способу получен функционал меньше, чем в предыдущем.

Способ поиска наименьшего элемента C_{ij} в матрице. Поставки распределяются, начиная с поиска наименьших величин C_{ij} в матрице:

$$10^1 - 15^2 - 25^2 - 10^2 - 20^3 - 5^4 - 5^5.$$

На основании распределения поставок получаем функционал:

$$F = (5 \cdot 5) + (25 \cdot 2) + (5 \cdot 4) + (15 \cdot 2) + (10 \cdot 1) + (20 \cdot 3) + (10 \cdot 2) = 215.$$

Сопоставляя величины функционала (F), полученные в результате составления базисного плана, получаем вывод: наименьший функционал (215), а значит, и наиболее оптимальное первоначальное распределение поставок получено в нашем примере по способу наименьшего элемента в матрице и строке.

b_j / a_i	B_1 40	B_2 25	B_3 15	B_4 10
A_1 30	⑤ 5	② 25	3	4
A_2 30	④ 5	2	② 15	① 10
A_3 20	③ 20	3	5	2
A_4 10	② 10	4	6	3

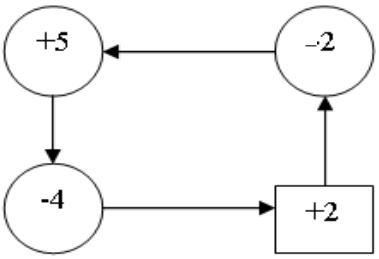
Во всех случаях распределения поставок клеток с кружками в матрицах меньше или равно $m + n - 1$, комбинации кружков вычеркиваемые, поэтому распределение поставок выполнено по установленным правилам.

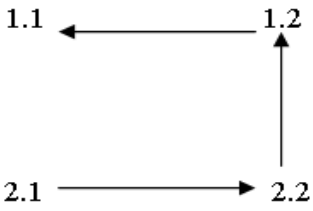
Изменение базисного допустимого плана. Для получения оптимального плана транспортной задачи следует выполнить условие минимизации F:

$$F = \sum_{i=1}^{m} \sum_{j=1}^{n} (C_{ij} X_{ij}) \to \min$$

путем изменения базисного допустимого плана. Для этого перемещаем меньшую поставку с большим C_{ij} в кружке в клетку, где нет поставки, а значение C_{ij} без кружка меньшее. Произведем перемещения в предыдущей матрице, составленной по способу наименьшего C_{ij} в строке.

Для реализации правила цепи, по которой должна перемещаться поставка, переместим поставку в клетке 2.1, равную 5, в клетку 2.2, где нет поставки. Перемещение проводим в направлении клеток, где есть поставки, и там же делаем повороты под прямым углом, пока цепь не замкнется. В нашем случае цепь имеет форму:





При перемещении поставки в вершинах цепи должны чередоваться плюсы и минусы. В клетке 2.2, куда вносим поставку, должен быть плюс, и ее обозначаем квадратом. Алгебраическая сумма C_{ij} по перемещаемым клеткам дает представление об увеличении функционала при получении положительной суммы или уменьшении – при получении отрицательной: (+5) + (–2) + (+2) + (–4) = +1. При указанном перемещении поставки базисный допустимый план ухудшился, так как алгебраическая сумма равна +1.

Других вариантов перемещения поставки по цепи в матрице произвести не можем, так как придется перемещать большие поставки из клеток с меньшей C_{ij} в клетки с большей C_{ij}, т. е. увеличивать функционал.

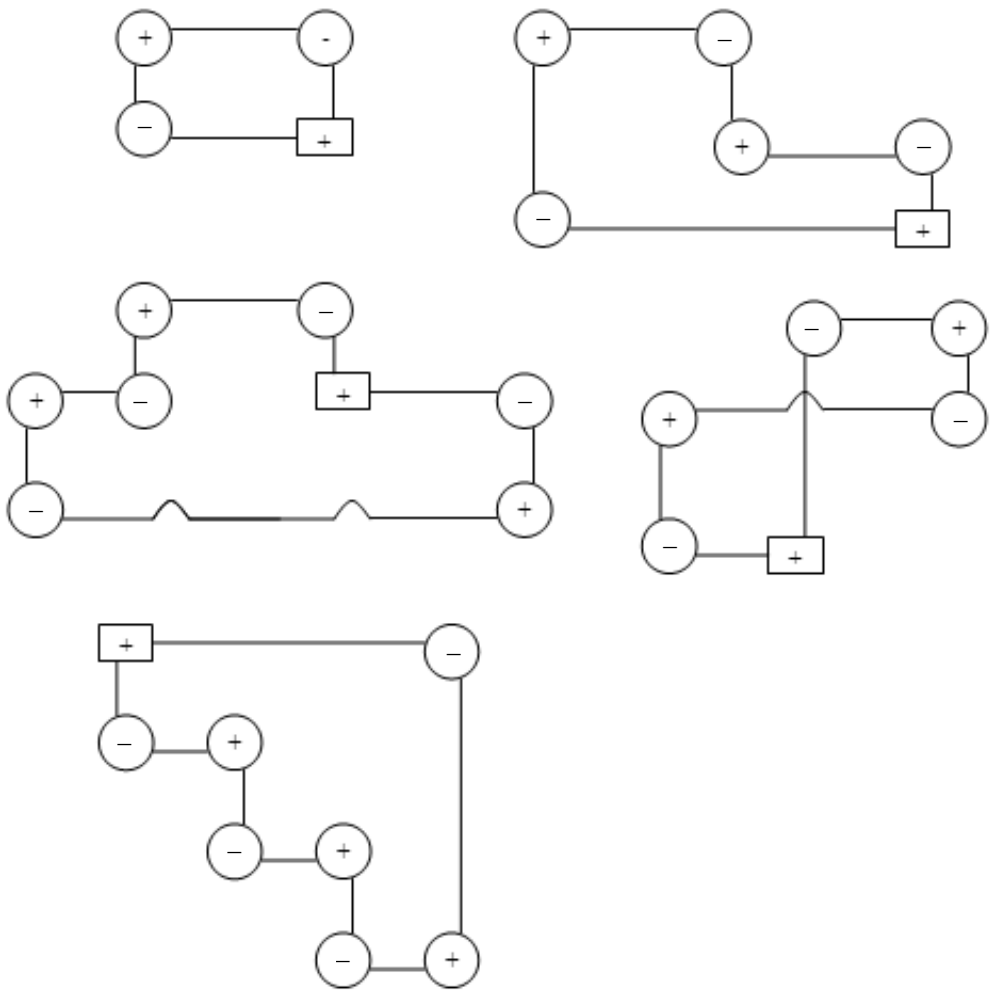
Если бы по условиям задачи необходимо было свести функционал к максимуму $F = \sum_{i=1}^{m}\sum_{j=1}^{n} (C_{ij}X_{ij}) \to max$, то такие перемещения улучшили бы функционал.

В клетках, где имеются поставки, при прохождении поставки по цепи их величина увеличивается (+) или уменьшается (–) на величину перемещаемой поставки (в нашем случае + или – 5).

После неудачного перемещения матрица примет вид:

b_j / a_i	B_1 40	B_2 25	B_3 15	B_4 10
A_1 30	⑤ 10	② 20	3	4
A_2 30	4	② 5	② 15	① 10
A_3 20	③ 20	3	5	2
A_4 10	② 10	4	6	3

Замкнутая цепь может иметь различную прямоугольную форму:



Правила построения цепи:

● Цепь должна быть замкнутым многоугольником.

● В цепи четное число вершин.

● Все углы цепи прямые.

● Отрезки цепи могут проходить через клетки матрицы, не являющиеся вершинами данной цепи, хотя в них могут содержаться поставки.

● Положительными (плюсовыми) вершинами будут клетки, в которых при перераспределении по цепи поставки увеличиваются.

● Отрицательными (минусовыми) вершинами являются клетки, в которых поставки при перераспределении уменьшаются.

● В цепи положительные вершины чередуются с отрицательными, количество их равно между собой.

● Вершина-квадрат, куда вносится поставка в ходе перераспределения, всегда положительная.

● При перераспределении поставок по цепи можно двигаться только по горизонтали или вертикали, изменяя направление только в вершинах цепи.

● Клетки, пересекаемые отрезками цепи, вершинами не являются, но в цепи отражаются в виде изломанной линии.

● Алгебраическая сумма чисел в вершинах цепи C_{ij} (–2 + 3 – 4 + 1 = –2) показывает, насколько может измениться значение функционала, если внести в вершину-квадрат поставку, равную 1. Сумму называют *характеристикой цепи*. Минусовая сумма указывает на уменьшение величины функционала, плюсовая – на увеличение. Эта величина равна произведению характеристики цепи (–2) на величину поставки (X_{ij}), которую мы перемещаем по цепи (5), т. е. $–2 \cdot 5 = –10$.

Вырождение матрицы – случаи в математике, которые являются исключением из общего правила. В транспортной задаче вырождение бывает в случаях, когда в допустимом плане поставок число клеток с кружками может быть меньше или больше, чем $m + n – 1$ (см. матрицу составления плана по способу северо-западного угла, где число клеток с кружками $< m + n – 1$, т. е. 6. В данном случае для решения проблемы вырождения в свободную клетку записывают нулевую поставку в порядке вычеркиваемой комбинации (клетка 2.4).

Если число кружков или число поставок в клетках больше $m + n – 1$, как в матрице *а*, необходимо выбрать наименьшую поставку (5) в клетке 3.2 и перераспределить ее по цепи, как в матрице *б* (см. ниже).

а)

a_i \ b_j	45	25	15	20
30			15	15
50	30	20		5
20	15	5		

Равенство мощностей и спросов (закрытая задача) создает потенциальную возможность вырождения, но не всегда приводит к нему.

б)

a_i \ b_j	45	25	15	20
30			15	15
50	25	25		5
20	20			

7.2. МЕТОД ПОТЕНЦИАЛОВ

Метод разработан в 1940 г. академиком Л. В. Конторовичем. В 1951 г. американским ученым Дж. Б. Данцигом предложен распределительный метод (МОДИ), аналогичный методу потенциалов. В обоих методах при проверке допустимого плана на оптимальность определяются *потенциалы* (числа), с помощью которых вычисляются *характеристики клеток без кружков* (в них нет поставок).

Обозначив потенциалы строк через U_i, потенциалы столбцов через V_j, показатели C_{ij} в клетках с поставками и кружками через $\overline{C_{ij}}$, характеристики клеток без кружков (без поставок) через E_{ij}, получим следующие соотношения:

метод потенциалов	метод МОДИ
$u_i = v_j - \overline{c_{ij}}$;	$u_i = \overline{c_{ij}} - v_j$;
$v_j = \overline{c_{ij}} + u_i$;	$v_j = \overline{c_{ij}} - u_i$;
$\overline{c_{ij}} = v_j - u_i$;	$\overline{c_{ij}} = v_j + u_i$;
$E_{ij} = c_{ij} - (v_j - u_i)$.	$E_{ij} = c_{ij} - (v_j + u_i)$.

Каждый показатель $\overline{c_{ij}}$ (в клетке матрицы он находится в кружке) должен быть равен разнице потенциалов своих столбцов и строк. Определение потенциала можно начинать с любой строки или столбца. Первый потенциал по величине выбирается произвольно (лучше определение начинать с нуля). Величины других потенциалов определяются с использованием предложенных выше формул, используя при первом вычислении выбранный нами потенциал.

Рассмотрим пример решения транспортной задачи, предложенный В. С. Михеевой (1981). Базисный допустимый план составлен способом наименьшего элемента в столбце, его первоначальный функционал 555 (табл. 7.4).

Вначале рассчитаем потенциалы строк и столбцов по методу потенциалов с использованием формул. Произвольно выбранную величину потенциала выбираем в том столбце или строке, где наибольшее количество клеток с кружками. В нашем примере – это третья строка. В качестве потенциала для нее возьмем число 0. По формуле (7.11) определяем потенциалы (v_j) первого ($v_j = \overline{c_{ij}} + u_i = 0 + 2 = 2$), третьего (0 + 6 = 6), четвертого (0 + 4 = 4), пятого (0 + 2 = 2) столбцов. Зная потенциалы по четырем столбцам, можно вычислить потенциалы строк (u_i) по формуле (7.10): первой ($u_i = v_j - \overline{c_{ij}} = 2 - 1 = 1$), четвертой (6 – 2 = 4). По полученным потенциалам новых строк вычисляем потенциалы новых столбцов, а по ним – новых строк (см. табл. 7.4).

Таблица 7.4

Базисный допустимый план (матрица 1)

a_i \ b_j	50	85	35	25	20	u_i
30	① **30**	2 *–4*	3 *–2*	4 *1*	5 *4*	1
40	4 *6*	③ **40**	2 *0*	3 *3*	4 *6*	4
70	② **20**	5 *–2*	⑥ **5**	④ **25**	② **20**	0
75	4 *6*	③ **45**	② **30**	1 *1*	1 *3*	4
v_j	2	7	6	4	2	

Определив потенциалы строк и столбцов, вычисляют характеристики клеток (E_{ij}) без кружков (в них нет поставок) по формуле (7.13). Приведем расчет характеристики клетки 1.2: $E_{ij} = c_{ij} – (v_j – u_i) = 2 – (7 – 1) = –4$. Аналогично рассчитываем E_{ij} (показаны курсивом) для других клеток без кружков.

Среди вычисленных характеристик клеток (курсив в клетках) отрицательные величины получены в клетках матрицы 1.2, 1.3, 3.2 (их величины соответственно: –4, –2, –2), поэтому план не оптимален.

Выбираем клетку с наибольшей отрицательной абсолютной величиной характеристики (E_{12}), равную –4. К клетке 1.2 строится цепь по перемещению наименьшей поставки 5 из клетки 3.3, так как F стремится к минимуму. Путь перемещения следующий:

3.3 (–5)→4.3 (+5)→4.2 (–5)→1.2 (+5)→1.1 (–5)→3.1 (+5)→3.3.

К имеющейся поставке в клетке прибавляется или отнимается 5 для сохранения баланса между поставщиками и потребителями.

Получаем матрицу с измененными поставками (7.5). В ней повторяем алгоритм расчетов, как в табл. 7.4: рассчитываем потенциалы строк и столбцов, характеристику клеток без поставок, производим перераспределение поставок с использованием другой минимальной поставки 25 в клетке 3.4. Другая минимальная поставка 25 в клетке 3.1 перемещаться

не может по цепи, так как в свободной клетке 4.4 с максимальной отрицательной абсолютной характеристикой (–3) ее следует вычитать из несуществующей поставки.

Таблица 7.5

Результаты первого перераспределения поставок (матрица 2)

a_i \ b_j	50	85	35	25	20	u_i
30	① **25**	② 5	3 *2*	4 *1*	5 *4*	1
40	4 *2*	③ **40**	2 *0*	3 *–1*	4 *2*	0
70	② **25**	5 *2*	⑥ *4*	④ **25**	② **20**	0
75	4 *2*	③ **40**	② **35**	1 *–3*	1 *–1*	0
v_j	2	3	2	4	2	

Поставка 25 в клетке 3.4 перемещается по цепи:

3.4 (–25)→3.1 (+25)→1.1 (–25)→1.2 (+25)→4.2 (–25)→4.4 (+25)→3.4.

Получена отрицательная характеристика –1 в клетке 4.5. Проводится новое перераспределение минимальной поставки равной 0 в клетке 1.1. Здесь нулевая поставка, так как из прежней поставки в клетке 25 следовало вычесть перераспределяемую 25. В таких ситуациях допускается наличие нулевой поставки, чтобы не нарушались правила перемещения поставки по цепи. Результаты распределения представлены в матрице 3 (табл. 7.6). В ней снизилось отрицательное значение E_{ij} до –1. План приблизился к оптимальному. Проводим очередное перераспределение поставок. Минимальную нулевую поставку из клетки 1.1 перемещаем в клетку 4.5, где отрицательная E_{ij}. Прибавление и вычитание нуля по цепи не изменяет величины поставок в клетках и не нарушает правил построения цепи.

Таблица 7.6

Результаты второго распределения поставок (матрица 3)

a_i \ b_j	50	85	35	25	20	u_i
30	① **0**	② 30	3 *2*	4 *4*	5 *4*	1
40	4 *2*	③ **40**	2 *0*	3 *2*	4 *2*	0
70	② **50**	5 *2*	6 *4*	④ *3*	② **20**	0
75	4 *2*	③ **15**	② **35**	1 **25**	1 *–1*	0
v_j	2	3	2	1	2	

Таблица 7.7

Результаты третьего распределения поставок (матрица 4)

a_i \ b_j	50	85	35	25	20	u_i
30	① **0**	② 30	3 *2*	4 *4*	5 *5*	1
40	4 *2*	③ **40**	2 *0*	3 *2*	4 *4*	0
70	② **50**	5 *2*	6 *4*	④ *3*	② **20**	–1
75	4 *2*	③ **15**	② **35**	1 **25**	① **0**	0
v_j	1	3	2	1	1	

В новой матрице 4 (табл. 7.7) после перерасчетов v_j, u_i, E_{ij} получены все *положительные* E_{ij} цепи при стремлении функционала к *минимуму*, поэтому план *оптимальный*, величина $F = 460$. По сравнению с базовым планом функционал снизился на 95 единиц.

7.3. ДЕЛЬТА-МЕТОД АГАНБЕГЯНА

Для решения закрытых и открытых транспортных задач А. Г. Аганбегян (1961) разработал дельта-метод. Он использовался для ручной обработки. Исходные данные используем из табл. 7.4. В каждом столбце этой таблицы находим минимальное значение c_{ij} и обводим его кружком. Если в столбце несколько равных по значению c_{ij}, выбираем любой из них (обычно первый сверху).

Вычисляем в каждом столбце приросты затрат (ΔC_{ij}) для строки как разницу между элементом c_{ij} строки и минимальным значением c_{ij} в столбце: $\Delta C_{ij} = c_{ij} - c_{ij\,\min}$. Для первого столбца значения ΔC_{ij} следующие (сверху вниз): 1 – 1 = 0; 4 – 1 = 3; 2 – 1 = 1; 4 – 1 = 3. Аналогично вычисляем ΔC_{ij} для других столбцов. В дальнейшем используем матрицу со значениями ΔC_{ij} в правом верхнем углу (табл. 7.8). В нее заносим поставки по столбцам, равные величине потребителя, в клетки с нулевыми значениями в кружках.

Таблица 7.8

Рабочая матрица прироста затрат

a_i \ b_j	50	85	35	25	20	d_i
30	0 **50**	0 **85**	1	3	4	*–105*
40	3	1	0 **35**	2	3	*5*
70	1 ↓	3	4	3	1	*70*
75	3	1	0	0 **25**	0 **20**	*30*

В дальнейшем производится расчет баланса (d_i), по которому определяем избыток или недостаток строк: $d_1 = 30 - (50 + 85) = -105$. Аналогично производится расчет баланса для последующих строк. Отрицательный баланс сложился лишь в первой строке. Значит план распределения поставок не оптимальный.

Производится перераспределение поставок из строк с минусовым балансом в строки с плюсовым балансом и учетом минимального значения прироста затрат (ΔC_{ij}). У нас минимальные значения $\Delta C_{ij} = 1$ в трех клет-

ках с положительным, но разным балансом в строках (3.1; 2.2; 4.2). Так как d_3 (70) больше, чем d_4 (30) и d_2 (5), выбирается клетка 3.1 для перемещения поставки (50) из соответствующего ей столбца 1. В строку с нулевым балансом поставка не перемещается.

В клетке 3.1 новой матрицы (табл. 7.9) обводим кружком ΔC_{ij}. В эту клетку (указано стрелкой) вносим поставку 50 из клетки 1.1, т. е. наименьшую из строки с отрицательным балансом –105. На величину 50 уменьшится отрицательный баланс первой строки и составит ($d_1 = 30 - 85 = -55$), а также положительный баланс третьей строки ($d_3 = 70 - 50 = 20$).

После первого перемещения поставки отрицательный баланс в первой строке сохранился. Необходимо продолжить перемещение поставки из минусовой строки в плюсовую по описанному выше алгоритму. Следует из клетки 1.2 переместить поставку в клетку 4.2 величиной не более d_4 (30). В результате перемещения новая матрица примет вид как в табл. 7.10.

Таблица 7.9

Первый вариант перемещения поставки

b_j / a_i	50	85	35	25	20	d_i
30	0	⓪ **85**	1	3	4	*–55*
40	3	1	⓪ **35**	2	3	*5*
70	① **50**	3	4	3	1	*20*
75	3	1 ↓	0	⓪ **25**	⓪ **20**	*30*

Второе перемещение поставки не привело к исчезновению отрицательного баланса первой строки ($d_1 = -25$). Следует переместить поставку из первой строки в клетке 1.2, равную этой величине d_1, в строку с положительным потенциалом. В табл. 7.10 положительный потенциал имеют вторая и третья строка. Их суммарная величина соответствует величине отрицательного баланса первой строки. Поэтому из поставки клетки 1.2 (55) сначала перемещаем 5 в клетку 2.2, так как прирост затрат здесь наименьший и новая матрица примет вид табл. 7.11.

Таблица 7.10

Второе перемещение поставки

a_i \ b_j	50	85	35	25	20	d_i
30	0	⓪ **55**	1	3	4	*–25*
40	3	1 ↓	⓪ **35**	2	3	*5*
70	① **50**	3	4	3	1	*20*
75	3	① **30**	0	⓪ **25**	⓪ **20**	*0*

Таблица 7.11

Третье перемещение поставки

a_i \ b_j	50	85	35	25	20	d_i
30	0	⓪ **50**	1	3	4	*–20*
40	3	① **5**	⓪ **35**	2	3	*0*
70	① **50**	3 ↓	4	3	1 ↑	*20*
75	3	① **30**	0 →	⓪ **25**	⓪ **20**	*0*

Затем переместим поставку 20 из клетки 1.2 в третью строку с положительным балансом 20 в клетку 3.5 с наименьшим приростом затрат ($\Delta C_{ij} = 1$). Оптимальный путь перемещения поставки 20 указан стрелками. В дельта-задаче цепь открытая. При перемещении поставки по цепи сохраняется чередование плюсов и минусов с изменением величин поставок на поворотах цепи под прямым углом как и в методе потенциалов. Новая матрица примет вид табл. 7.12, где нет отрицательного баланса, все значения его по строкам нулевые.

При перемещении поставки в другие клетки увеличивается прирост затрат. Функционал будет стремиться к максимуму вместо минимума. Следовательно, получен оптимальный план размещения поставок с использованием дельта-метода.

Таблица 7.12

Четвертое перемещение поставки

a_i \ b_j	50	85	35	25	20	d_i
30	0	⓪ **30**	1	3	4	*0*
40	3	① **5**	⓪ **35**	2	3	*0*
70	① **50**	3	4	3	① **20**	*0*
75	3	① **50**	0	⓪ **25**	⓪ **0**	*0*

Проверка решения. В дельта-методе поиск клетки в плюсовой строке, к которой следует строить цепь, не формализован и опирается на мыслительную способность человека. Поэтому могут быть допущены ошибки при выборе наиболее выгодных цепей, в результате этого будет получен допустимый план вместо оптимального.

В оптимальном варианте распределения поставок (табл. 7.12) можно рассчитать потенциалы строк и столбцов и характеристики клеток без кружков, чтобы убедиться в отсутствии отрицательных характеристик как в методе потенциалов, но несколько иным способом. Для той плюсовой строки, которая на последнем шаге вычислительных операций превратилась в нулевую, берется потенциал равный нулю. Для минусовой строки, которая на последнем шаге вычислительных операций превратилась также в нулевую, берется потенциал, равный приросту затрат на этом последнем шаге, т. е. алгебраическая сумма цепи: –0 (клетка 1.2) + 1 (4.2) + (–0 (4.5)) + 1 (3.5) = +2. В нашем примере третья строка имеет потенциал равный нулю, а первая с отрицательным балансом – +2. Расчет потенциалов остальных строк и столбцов осуществляется по формулам Конторовича (7.10–7.13). Для проверки правильности решения можно пользоваться матрицей со значениями $\overline{C_{ij}}$ или ΔC_{ij}.

Проще проверить оптимальность плана по дельта-методу вычислением функционала и сравнения его с функционалом табл. 7.7 ($F = 460$):

$$F = \sum(\overline{C_{ij}} \cdot X_{ij}) = 2 \cdot 50 + 2 \cdot 30 + 3 \cdot 5 + 3 \cdot 50 + 2 \cdot 35 + 1 \cdot 25 + 2 \cdot 20 = 460.$$

Величина $\overline{C_{ij}}$ взята для соответствующих клеток табл. 7.12 из табл. 7.7. Оба метода распределения поставок показали один и тот же результат функционала, равный 460.

Отличие построения цепей в дельта-методе:

– цепь строится незамкнутая;

– цепь начинается в клетке с кружком (с поставкой), которая находится в минусовой строке; в этой клетке поставка уменьшается, и она становится отрицательной вершиной цепи;

– перемещение поставки в конец открытой цепи производится как в методе потенциалов с чередованием положительных и отрицательных вершин;

– в этом методе не требуется количества кружков (клеток с поставками), равного $m + n - 1$;

– в исходном плане число кружков равно числу столбцов, лишь в ходе решения появляются новые клетки с кружками (поставками);

– в незамкнутой цепи вершинами бывают клетки без кружков (без поставок); они положительны, так как в них вносится поставка;

– характеристика незамкнутой цепи рассчитывается как алгебраическая сумма показателей ΔC_{ij} или $\overline{C_{ij}}$ в ее вершинах; так как при распределении поставок по цепи функционал увеличивается, характеристика цепи всегда положительная; она показывает, насколько увеличивается величина функционала, если передвинуть по цепи поставку, равную 1, из минусовой строки в плюсовую.

7.4. МОДИФИКАЦИЯ МОДЕЛЕЙ ТРАНСПОРТНЫХ ЗАДАЧ

Транспортные задачи могут быть открытыми (учитывать время транспортировки продукции, затраты на производство единицы продукции), многоэтапными, многопродуктовыми. Все они, как и закрытая транспортная задача, являются частным случаем более сложной лямбда-задачи. Рассмотрим некоторые из них.

7.4.1. Открытая транспортная задача

Транспортная задача, в которой суммарная мощность поставщиков не совпадает с суммарным спросом потребителей, называется открытой. В связи с этим условия модели записываются как: $\sum a_i > \sum b_j$ или $\sum a_i < \sum b_j$.

Для решения открытой транспортной задачи могут применяться методы: потенциалов, дельта-метод, МОДИ.

При решении задачи методом потенциалов или МОДИ проводятся следующие дополнительные мероприятия. Если суммарные мощности поставщиков превышают суммарные мощности потребителей, в матрицу исходных данных следует ввести дополнительный столбец – *фиктивный потребитель* (*B*) со спросом, равным небалансу: $b_{n+1} = \sum a_i - \sum b_j$. Показатели c_{ij} в столбце фиктивного потребителя должны быть *одинаковыми* по величине, которая устанавливается произвольно (любая величина, обычно проставляют 0).

Если суммарный спрос потребителей превышает суммарную мощность поставщиков, необходимо ввести в матрицу дополнительную строку – *фиктивного поставщика (A)*, мощность которого должна быть равна небалансу: $a_{m+1} = \sum b_j - \sum a_i$. Показатели c_{ij} этой строки должны быть *одинаковыми* и произвольными (обычно нулевые).

При составлении базисного допустимого плана и в процессе вычислительных операций в матрице должно содержаться число поставок (клеток с кружками), равное $m + n - 1$. Они должны находиться в порядке вычеркиваемой комбинации. Учитываются фиктивные строки и столбцы.

При использовании дельта-метода фиктивные поставщики или потребители не вводятся. Задача решается с нарушением баланса строк и столбцов.

7.4.2. Максимизация целевой функции

Многие экономгеографические задачи требуют максимизации функции (повышение производительности труда, прибыли и т. д.):

$$F = \sum_{i=1}^{m} \sum_{j=1}^{n} (C_{ij} X_{ij}) \rightarrow \max.$$

При использовании метода потенциалов, МОДИ в базисном допустимом плане поставки размещаются в клетках с *наибольшим значением* c_{ij}. Перераспределение поставок производится с учетом построения цепи к клеткам с *наибольшей положительной характеристикой*. Оптимальным будет такой план перераспределения поставок, в котором характеристики клеток без кружков будут отрицательными и нулевыми.

Решение задач дельта-методом следует начинать с распределения поставок в клетки, в которых показатели c_{ij} имеют максимальные величины.

Транспортная задача с максимизацией функции может решаться по способу ее минимизации при условии придания всем c_{ij} отрицательных значений. Получив оптимальный план, необходимо рассчитать значение

целевой функции, используя c_{ij} до их преобразования в отрицательные величины.

Решение транспортных задач может происходить при условии ограничения поставок или потребления: «не меньше, чем» (обязательные поставки) и «не больше, чем». Конечный результат решения таких задач не достигает оптимальных условий, поэтому их следует преобразовать в закрытую задачу.

7.4.3. Ограничения по времени транспортировки продукции

Ограничения в транспортную задачу вводят при учете *времени транспортировки продукции.* Для этого в искомом оптимальном плане не должны быть такие перевозки между поставщиками и потребителями, временная продолжительность которых больше заданной величины.

Пример. Требуется решить задачу на минимум совокупных затрат на транспортировку продукции. Длительность перевозки не может превышать 4 часов. В матрицу *a* (табл. 7.13) вводится дополнительно время перевозки продукции (f_{ij}) и размещается в левом верхнем углу клетки.

При максимальном времени перевозки продукции (4 часа) запрещаются поставки в клетки, где время указано больше: 1.1 (7 ч), 1.4 (5 ч), 3.2 (6 ч), 4.3 (5 ч). После этого задача может решаться любым методом. Ее итоговое решение (оптимальный план) представлен в табл. 7.13 *b*.

Иногда введение ограничений приводит к невозможности построить даже единственно допустимый план, который в таком случае был бы оптимальным. Значит, исходная информация противоречит условиям содержательной математической постановки задачи, которая по этой причине не имеет решения.

Таблица 7.13

Учет времени перевозки продукции

a

a_i \ b_j	**17**	**10**	**35**	**23**
30	7 3	2 3	2 2	5 1
20	3 2	2 1	3 2	2 2
15	1 2	6 2	4 3	2 4
20	2 1	4 1	5 3	1 2

b

a_i \ b_j	**17**	**10**	**35**	**23**
30	M	2 3	② **30**	M
20	3 2	2 1	3 2	② **20**
15	② **10**	M	③ **5**	2 4
20	① **7**	① **10**	M	② **3**

7.4.4. Транспортно-производственная задача

В географических исследованиях, посвященных вопросам определения границ зон сбыта продукции или рациональных связей по прикреплению потребителей к поставщикам, должны учитываться не только транспортные, но и производственные затраты. Такие задачи получили название транспортно-производственных. В качестве $c_{ij} = S_i + t_{ij}$ выступают транспортно-производственные затраты, т. е. S_i – затраты на производство единицы продукции (себестоимость, цена единицы продукции или приведенные удельные затраты) i-м поставщиком; t_{ij} – затраты на перевозку продукции между i-м поставщиком и j-м потребителем. Если увеличить или уменьшить на одну и ту же величину все показатели c_{ij} в матрице, или в строке, или в столбце, то свойства матрицы не изменятся. Суммарные мощности поставщиков равны суммарному спросу потребителей. Следовательно, какой бы ни была стоимость производства, потребители для удовлетворения своего спроса возьмут продукцию у всех поставщиков. От каких поставщиков каждый потребитель получит продукцию зависит от транспортных затрат.

Решение открытой транспортно-производственной задачи должно учитывать показатель S_i, например, себестоимость продукции. При суммарной мощности поставщиков, предположим на 20 единиц, превышающих суммарный спрос потребителей, у последних появляется свобода выбора в получении продукции от более выгодных поставщиков, поэтому оптимальный план может быть экономически более эффективным.

Модель транспортно-производственной задачи при введении дополнительных условий можно использовать для оптимизации развития и размещения промышленного производства, получить ответ, где должны располагаться новые промышленные объекты.

Для решения этих закрытых и открытых задач используются рассмотренные методы функционала, потенциала.

7.4.5. Многоэтапная транспортная задача

В современных условиях перевозка продукции от поставщика к потребителю осуществляется двумя путями: *поставщик → потребитель* (наиболее экономически выгодный) и *поставщик → база → потребитель* (требует больше транспортных и иных затрат). Поставка продукции через базу к потребителю требует построения модели многоэтапной транспортной задачи, в которой за критерий оптимальности обычно принимается минимальное значение совокупных транспортных затрат. Способ

решения транспортных задач с двумя и более этапами предложен американским ученым А. Орденом. Впоследствии его назвали *способом фиктивной диагонали*.

План перевозки между поставщиками и складами и план перевозки между складами и потребителями не зависят друг от друга. Решаются две самостоятельные транспортные задачи раздельно и в любом порядке.

Если суммарная мощность складов больше суммарной мощности поставщиков, то необходимо осуществлять единый расчет, чтобы получить экономически более эффективный план многоэтапных перевозок. Рассмотрим построение матрицы в двухэтапной задаче (табл. 7.14) при следующих условиях:

$$\sum D_p > \sum A_i, \ \sum A_i = \sum B_j.$$

Таблица 7.14

Форма записи исходных данных в четырехблочную матрицу

Поставщики и их мощности	Потребители и их спрос						
	D_1 50	D_2 *50*	D_3 *50*	B_1 *20*	B_2 *25*	B_3 *15*	B_4 *30*
A_1 55	7	5	4	M	M	M	M
A_2 35	2	3	4	M	M	M	M
D_1 50	0	M	M	7	5	3	5
D_2 50	M	0	M	3	4	5	6
D_3 50	M	M	0	10	9	8	7
	III			IV			

При различных возможных вариантах использования емкостей складов другими могут быть варианты перевозок грузов между складами и потребителями. В матрице (см. рис. 7.14) в вектор поставщиков попадают истинные поставщики (A_i) и склады (D_p), так как склады выступают по отношению к истинным (конечным) потребителям (B_j) как поставщики. В вектор потребителей попадают истинные потребители и склады, получающие продукцию от поставщиков. По этой причине матрица состоит из четырех блоков.

Элементами первого (I) блока (левого верхнего прямоугольника) являются затраты на перевозку грузов между поставщиками и складами. Во втором блоке (II – правом верхнем прямоугольнике) все клетки содержат запреты (M), так как поставщики передают свою продукцию сначала на склад, прямых связей с потребителями не имеют. Элементами четвертого (IV) блока (правого нижнего прямоугольника) являются за-

траты на перевозку грузов от складов к потребителям. В третьем (III) блоке (левом нижнем прямоугольнике) склады не поставляют продукцию складам, поэтому во всех клетках, за исключением диагональных, проставляются запреты (М). Запись поставок в фиктивную диагональ будет символизировать недоиспользованную емкость складов.

Фиктивная диагональ вводится для того, чтобы связать I и IV блоки. Суммарный размер поставок в блоках I и III по каждому столбцу равен емкости соответствующего склада. Суммарный размер поставок в блоках III и IV по каждой строке равен емкости склада.

Решение задач по блочным матрицам не отличается от алгоритма транспортных задач. Имеются лишь различия в составлении базисного плана. Его построение надо начинать с распределения поставок в одном из двух блоков I или IV. Затем следует определить, где осталась неиспользованная часть емкости складов, записать «поставки» в соответствующие клетки фиктивной диагонали. С учетом этих «поставок» можно переходить к построению плана распределения поставок в оставшийся блок, IV или I. Требование к числу кружков, равному $m + n - 1$, расположенных в порядке вычеркиваемой комбинации, предъявляется к матрице в целом.

7.4.6. Многопродуктовая транспортная задача

Все рассмотренные транспортные задачи относятся к числу однопродуктовых. Однако иногда возникает необходимость составления базисного плана перевозок взаимозаменяемых видов продукции. Такой вопрос следует решать как единую задачу, так как в ней различные продукты могут приравниваться друг к другу через переводные коэффициенты. Решение задачи данной модели не имеет принципиальных отличий от решения закрытой однопродуктовой задачи. Существуют лишь специфические методические приемы обработки исходной информации, которые необходимо знать, чтобы подготовить матрицу для выполнения расчетов.

Пример. Потребителю необходимо поставить взаимозаменяемое топливо: торф, бурый уголь. Необходимое условие: суммарная потребность в торфе и буром угле, выраженная в единицах условного топлива, будет полностью удовлетворена. Известно, что 1 т условного топлива равна 7000 ккал, 1 т торфа – 2800 ккал, 1 т бурого угля – 4200 ккал. Отсюда переводной коэффициент по теплотворной способности топлива (калорийный эквивалент) для торфа равен 2800 / 7000 = 0,4, для бурого угля – 0,6.

В табл. 7.15 представлены мощности и спросы по торфу в тоннах и показан оптимальный план перевозки с функционалом, равным 13 980 (F_1),

в табл. 6.16 представлены эти же данные по бурому углю с функционалом 10 620 (F_2). По двум планам объем грузооборота равен $F_1 + F_2 =$ = 24 600 т/км. У поставщиков A_1 и A_2 имеется торф и бурый уголь, у поставщика A_3 – только торф, у поставщика A_4 – только бурый уголь. В обеих таблицах расстояния между поставщиками A_1 и A_2 и потребителями одинаковые, так как оба вида топлива будут перевозиться по одним и тем же транспортным путям.

Используя коэффициенты теплотворной способности торфа (0,4) и бурого угля (0,6), данные A_i, B_j, x_{ij}, c_{ij} таблиц 7.15, 7.16, производим перерасчет и составляем табл. 7.17, в которой данные указаны в условных (перерасчетных) единицах. Приводим ниже пояснения, связанные с перерасчетом.

1. Расчет спроса потребителей в условных единицах (у. е.) проведем на примере B_1 (см. табл. 7.15, 7.16). Спрос потребителя B_1 на торф равен 100 т, на бурый уголь – 60 т. Используя переводные коэффициенты, рассчитываем его потребность в условном топливе:

$$B_1 = 100 \cdot 0{,}4 + 60 \cdot 0{,}6 = 76.$$

Таблица 7.15

Мощности и спросы по торфу

b_j / a_i	B_1 100	B_2 180	B_3 120
A_1 150	**12** **100**	72	**60** **50**
A_2 75	48	**24** **5**	**48** **70**
A_3 175	72	**36** **175**	60

Таблица 7.16

Мощности и спросы по бурому углю

b_j / a_i	B_1 60	B_2 210	B_3 125
A_1 130	**12** **60**	72	**60** **70**
A_2 100	48	**24** **45**	**48** **55**
A_4 165	36	**12** **165**	72

2. Мощность поставщиков (A_i) в условных единицах дается отдельно для каждого вида топлива, потому что она выступает как ограничение на возможный размер поставок *k*-го вида продукта, который находится у *i*-го поставщика. Ее получают путем умножения величины мощности поставщика на переводной коэффициент по торфу и по бурому углю отдельно:

$A_1 = 150 \cdot 0{,}4 = 60$ (по торфу); $A_1 = 130 \cdot 0{,}6 = 78$ (по бурому углю).

3. Показатели расстояний (c_{ij} – *правый верхний угол в клетках матриц, курсив, полужирный шрифт*) в условных единицах получают путем деления их на переводные коэффициенты:

$c_{11} = 12 / 0{,}4 = 30$ (по торфу); $c_{11} = 12 / 0{,}6 = 20$ (по бурому углю).

После перевода всех показателей матрицы в условные единицы, как показано в пунктах 1–3, получаем новую матрицу, в которой проводим перераспределение условных единиц до получения оптимального варианта (табл. 7.17).

Таблица 7.17

Оптимальный вариант распределения поставок в условных единицах

b_j, у. е. / a_i, у. е.			B_1 76	B_2 198	B_3 123
A_1	торф	60	**30** **60**	150	120
	бурый уголь	78	**20** **16**	100	**80** **62**
A_2	торф	30	150	**60** **29**	**90** **1**
	бурый уголь	60	100	40	**60** **60**
A_3	торф	70	180	**90** **70**	150
A_4	бурый уголь	99	60	**20** **99**	100

Функционал оптимального плана поставок, выраженный в условных единицах в табл. 7.17, составляет 20 790. По сравнению с суммарным потенциалом предыдущих таблиц по торфу и бурому углю, объем транспортной работы в последнем варианте с условными единицами снизился на 3810 единиц, что дает экономию по объему грузооборота более, чем на 15 %.

Цель достигнута, задача решена. Получен оптимальный план перевозки топлива.

7.4.7. Лямбда-задача

Алгоритм транспортных задач по методам решения значительно проще, чем лямбда-задачи. Ее называют распределительная или обобщенная транспортная задача. В ее модели отражается более широкий круг практических задач богатых по содержанию. Способ ее решения предложили американские математики А. Фергюссон, Дж. Данциг (1955). Позже ее разрабатывали российские ученые А. Л. Брудно, У. Х. Малков, А. Г. Аганбегян и др.

Алгоритм У. Х. Малкова строго формализован и реализован в машинных программах, дает возможность преодолеть трудности решения лямбда-задач, но очень сложный. А. Г. Аганбегян предложил операторскую схему дельта-метода решения лямбда-задач, которую можно использовать для расчетов вручную. Основные принципы этого метода изложены выше в 8.5 применительно к транспортной задаче. Решение лямбда-задачи дельта-методом также сложно. В ходе вычислительных операций возможно частое допущение ошибок, поэтому итоговое решение следует проверять. Лямбда-задача открытая, в ходе ее решения всегда останется хотя бы одна избыточная (плюсовая) строка. Потенциал (оценка) этой строки принимается равным нулю, а расчет потенциалов других строк и столбцов, характеристик клеток без кружков осуществляется по следующим формулам:

$$u_i = \overline{\lambda_{ij}}\ (v_j - \overline{c_{ij}});$$

$$v_j = \overline{c_{ij}} + u_i / \overline{\lambda_{ij}};$$

$$\overline{c_{ij}} = v_j - u_i / \overline{\lambda_{ij}};$$

$$E_{ij} = c_{ij} - (v_j - u_i / \overline{\lambda_{ij}}).$$

Показатель $\overline{\lambda_{ij}}$ размещается в левом верхнем углу клеток матрицы и выполняет важную роль в оптимизации условий задачи, включается в формулу при расчете функционала: $F = \sum (c_{ij} \cdot \overline{\lambda_{ij}} \cdot x_{ij}) \rightarrow \min (\max)$.

Для той минусовой строки, которая на последнем шаге вычислительных операций превратилась в нулевую, рекомендуется взять потенциал, равный приросту затрат на последнем шаге.

Пример полного решения лямбда-задачи приведен в курсе лекций В. С. Михеевой (1981, с. 107–154).

7.4.8. Оптимизация трансформации сельскохозяйственных угодий

При внутрихозяйственном землеустройстве проводится трансформация сельскохозяйственных угодий, т. е. перевод угодий из одного вида в другой. Это необходимо в связи с возникшими новыми производственными задачами, при повышении удельного веса ценных сельскохозяйственных угодий, укрупнении земельных массивов путем освоения новых земель и комиссации угодий, ликвидации мелкоконтурности участков, улучшения их культуртехнического состояния, изменении специализации хозяйства, при защите почв от эрозии и др.

При наличии ограниченных ресурсов, отпускаемых на трансформацию угодий, необходимо найти такой план, который обеспечит хозяйству получение максимального экономического эффекта.

Математическая модель задачи формируется следующим образом: в качестве неизвестных (x_{ij}) выступает площадь i-го угодья, трансформируемого в j-е, а также площади объектов мелиорации, имеющие в составе различные угодья.

В модель вводятся ограничения.

1. Наличие пригодных для трансформации земель:

$$\sum_j x_{ij} \le P_i, i \in M_1,$$

где P_i ($i \in M_1$) – площадь угодий пригодная для трансформации, га.

2. Затраты денежных средств на трансформацию:

$$\sum_j a_{ij} x_{ij} \le A_i, i \in M_2,$$

где a_{ij} – затраты денежных средств на перевод угодья из одного вида в другой, руб. га; A_i – объем ежегодных производственных затрат на осуществление трансформации угодий, руб.

3. Трудовые ресурсы:

$$\sum_j t_{ij} x_{ij} \le T_i, i \in M_3,$$

где t_{ij} – затраты труда на перевод единицы i-го угодья в j-е, чел.-дн. на 1 га; T_i – объем трудовых ресурсов на трансформацию в i-й период, чел.-дн.

4. Наличие машин и механизмов:

$$\sum_j l_{ij} x_{ij} \le L_i, i \in M_4,$$

где l_{ij} – норма затрат механизированных ресурсов на перевод единицы i-го угодья в j-е, усл. эт. га; L_i – j, объем работ i-го вида, выполняемых машинами, усл. эт. га.

5. Потребности в удобрениях:

$$\sum_j w_{ij} x_{ij} \le W_i, i \in M_5,$$

где w_{ij} – дозы вносимых удобрений в трансформируемые угодья, ц усл. ед; W_i – количество имеющихся удобрений i-го вида, ц.

6. Капиталовложения, выделяемые на трансформацию:

$$\sum_j d_{ij} x_{ij} \le D_i, i \in M_6,$$

где d_{ij} – норма затрат капиталовложений на перевод угодья из i-го вида в j-й, руб.; D_i – общий объем капиталовложений, расходуемых на трансформацию, руб.

Аналогично могут быть построены ограничения по другим ресурсам. Срок окупаемости капитальных вложений (T^t) рассчитывается по формуле:

$$T^t = \frac{\sum_{ij} d_{ij} x_{ij}}{\sum_{ij} g_{ij} x_{ij}},$$

где g_{ij} – дополнительный чистый доход, получаемый при переводе i-го вида угодья в j-е, руб.

Величина, обратная сроку окупаемости капитальных вложений, называется коэффициентом эффективности капитальных вложений (E):

$$E = 1 / T^t.$$

Чем больше коэффициент, тем меньше срок окупаемости затрат. При решении задач принимают $E = E_н$. Величину $E_н$ устанавливают, исходя из принимаемого срока окупаемости затрат капиталовложений. Если T^t принимается равным 5 лет, то $E_н = 0{,}2$; если 10 лет, то $E_н = 0{,}1$.

Целевая функция решаемых задач имеет вид:

$$F = \sum_{ij} (c_{ij} x_{ij}) \to \max.$$

В качестве c_{ij} можно использовать чистый доход g_i или прирост чистого дохода g_{ij}. Прирост чистого дохода рассчитывают:

$$g_{ij} = g_j - g_i,$$

где g_j – чистый доход после трансформации угодий; g_i – чистый доход до трансформации угодий.

Чистый доход рассчитывают по формулам:

$$g_i = B_i - H_j,\ g_i = B_j - H_i,$$

где B_i и B_j – стоимость валовой продукции соответственно до и после трансформации, руб.; H_i и H_j – себестоимость продукции соответственно до и после трансформации, руб.

Для решения задачи собирают исходную информацию, определяют состав переменных, рассчитывают показатели, необходимые для составления модели.

Исходная информация. Затраты денежных средств на перевод угодья из одного вида в другой, руб./га; затраты труда, чел.-дн./га; объем механизированных работ, эт. га; дозы удобрений; планируемая и фактическая урожайность; продуктивность угодий; себестоимость продукции (фактическая и планируемая); закупочные цены; площади земель, пригодные для трансформации, га; объемы работ, выполняемых имеющейся и поставляемой техникой.

Расчетные показатели. Дополнительный чистый доход при переводе одного вида угодий в другой; чистый доход с 1 га угодья до и после освоения.

Пример. В сельскохозяйственном предприятии выделено четыре поля, пригодных для трансформации в другие виды угодий и улучшения их. Намечено шесть видов их использования и определены шесть неизвестных переменных x_{ij} (табл. 7.18): X_1 – сад, трансформированный из пашни; X_2 – пашня, трансформированная из сенокосов; X_3 – сенокос, улучшенный, трансформированный из бывшего сенокоса; X_4 – пашня, трансформированная из пастбища; X_5 – пастбище, улучшенное из бывшего пастбища; X_6 – пастбище, улучшенное, трансформированное из прочих земель.

Таблица 7.18

Переменные при трансформации угодий

Угодья по проекту / Угодья на год землеустройства	Сады	Пашня	Сенокосы, улучшенные	Пастбища, улучшенные	Площадь, пригодная для трансформации
Пашня	X_1				200
Сенокосы		X_2	X_3		400
Пастбища		X_4		X_5	600
Прочие				X_6	200

Капиталовложения на трансформацию составляют 200 млн руб. объем трудовых ресурсов 8000 чел.-дн. Необходимо составить план трансформации, который обеспечит хозяйству максимальную экономическую эффективность с учетом денежных средств и трудовых ресурсов.

Экономико-математическая модель. В качестве целевой функции используем максимальный чистый доход после трансформации. Используя данные табл. 7.19, рассчитаем значения c_{ij} для модели целевой функции:

$$c_1 = 40 \cdot 50 - 800 = 1200;$$
$$c_2 = 30 \cdot 10 - 180;$$
$$c_3 = 50 \cdot 3 = 110;$$
$$c_4 = 30 \cdot 10 - 120 = 180;$$
$$c_5 = 80 \cdot 0{,}9 - 30 = 42;$$
$$c_6 = 80 \cdot 0{,}9 - 30 = 42.$$

Подставим полученные значения c_{ij} в уравнение целевой функции:

$$F = \sum (c_{ij}\, x_{ij}) \rightarrow \max = 1200\, x_1 + 180\, x_2 + 110\, x_3 + 180\, x_4 + 42\, x_5 + 42\, x_6.$$

На неизвестные накладываются следующие ограничения:

1. По площади: $x_1 \le 200$; $x_2 + x_3 \le 400$; $x_4 + x_5 \le 600$; $x_6 \le 200$.
2. По капитальным вложениям:

$$300\, x_1 + 100\, x_2 + 50\, x_3 + 80\, x_4 + 50\ x_5 + 800\, x_6 = 200\,000.$$

3. По трудовым ресурсам:

$$20\, x_1 + 2\, x_2 + 1{,}50\, x_3 + 2\, x_4 + 1{,}5\ x_5 + 30\, x_6 \le 8000.$$

4. По эффективности капитальных вложений и трансформацию. До составления ограничения необходимо рассчитать коэффициенты g_{ij}, показывающие прирост чистого дохода: $g_i = B_j - H_i$:

$$g_1 = (40 \cdot 50 - 800) - (20 \cdot 10 - 100) = 1100;$$
$$g_2 = (30 \cdot 10 - 120) - (20 \cdot 3 - 30) = 150;$$
$$g_3 = (50 \cdot 3 - 40) - (20 \cdot 3 - 30) = 80;$$
$$g_4 = (30 \cdot 10 - 120) - (40 \cdot 0{,}9 - 27) = 171;$$
$$g_5 = (80 \cdot 0{,}9 - 30) - (40 \cdot 0{,}9 - 27) = 33;$$
$$g_6 = (80 \cdot 0{,}9 - 30) - 0 = 42.$$

Рассчитываются общие коэффициенты (K_{ij}) при x_{ij}, приняв коэффициент эффективности капиталовложений ($E_н$) равным 0,1:

при $x_1 - K_1 = D_1 \cdot E_н - g_1 = 300 \cdot 0{,}1 - 1100 = -1070$;
при $x_2 - K_2 = D_2 \cdot E_н - g_2 = 100 \cdot 0{,}1 - 150 = -140$;
при $x_3 - K_3 = D_3 \cdot E_н - g_3 = 50 \cdot 0{,}1 - 80 = -75$;

Таблица 7.19

Исходные данные для расчета технико-экономических коэффициентов

Угодья на год землеустройства	Намечаемое использование	Переменные	Затраты на трансформацию		Урожайность, ц/га		Стоимость единицы продукции, руб.		Производственные затраты, руб./га	
			капиталовложения, руб./га	трудовые ресурсы, чел.-дн./га	до трансформации	после трансформации	до трансформации	после трансформации	до трансформации	после трансформации
Пашня	Сад	X_1	300	20	20	40	10	50	100	800
Сенокосы	Пашня	X	100	2	20	30	3	10	30	120
	Сенокосы, улучшенные	X_3	50	1,5	20	50	3	3	30	40
Пастбища	Пашня	X_4	80	2	40	30	0,9	10	27	120
	Пастбища, улучшенные	X_5	50	1,5	40	80	0,9	0,9	27	30
Прочие	Пастбища, улучшенные	X_6	800	30	0	80	0	0,9	0	30

при $x_4 - K_4 = D_4 \cdot E_н - g_4 = 80 \cdot 0{,}1 - 171 = -163$;
при $x_5 - K_5 = D_5 \cdot E_н - g_5 = 50 \cdot 0{,}1 - 33 = -28$;
при $x_6 - K_6 = D_6 \cdot E_н - g_6 = 800 \cdot 0{,}1 - 42 = 38$.

Таким образом, ограничение по эффективности капиталовложений примет вид: $-1070\,x_1 - 140\,x_2 - 75\,x_3 - 164\,x_4 - 28\,x_5 + 38\,x_6 \leq 0$.

После соответствующих расчетов по программе ЭВМ получены следующие результаты. В хозяйстве необходимо в ходе трансформации угодий:

1. Заложить сад на площади 200 га на пашне ($x_1 = 200$).
2. Освоить $x_2 = 400$ га сенокосов под пашню.
3. Трансформировать 600 га (x_4) пастбищ в пашню.
4. Перевести 65 га (x_6) прочих угодий в улучшенные пастбища.

В результате трансформации хозяйство получит максимальный чистый доход (F_{max}), равным 422 730 руб.

7.4.9. Модель формирования сырьевых зон перерабатывающих предприятий

Для математической модели формирования сырьевых предприятий используется распределительный метод линейного программирования (функционала или потенциала). Сырьевая культура должна размещаться в хозяйствах, где возможно получение наибольшего чистого дохода или необходим минимум затрат на ее производство. Чтобы уменьшить транспортные расходы, посевы размещают вблизи перерабатывающих предприятий. Снижение затрат на переработку сырья определяется мощностью предприятий, которые обеспечивали бы их минимумом продукции. Указанные критерии производства учитывают эффективность его в отдельном структурном звене. Для всех структурных звеньев аграрного промышленного комплекса следует рассчитать комплексный показатель – *минимум удельных приведенных затрат* всех звеньев в качестве критерия оптимальности (ПЗу) (по Е. М. Чепурному, 2001):

$$ПЗ_y = \frac{ПЗ_{сх} + ПЗ_и + ПЗ_{пп}}{V} = \frac{(С_{сх} + Е_н К_{сх}) + (С_и + Е_н К_и) + (С_{пп} + Е_н К_{пп})}{V},$$

где $ПЗ_y$ – удельные приведенные затраты, руб. за 1 т; $ПЗ_{сх}$, $ПЗ_и$, $ПЗ_{пп}$ – приведенные затраты соответственно в сельскохозяйственном производстве, инфраструктуре и перерабатывающей отрасли, тыс. руб.; $С_{сх}$, $К_{сх}$ – издержки и капитальные вложения на сельскохозяйственных предприятиях по производству сырья, тыс. руб.; $С_и$, $К_и$ – издержки и капитальные вложения в инфраструктуре отраслевого подкомплекса регионального АПК,

связанные с доставкой и хранением сырья, тыс. руб.; $С_{пп}$, $К_{пп}$ – соответственно издержки на переработку сырья в конечную продукцию и капитальные вложения на перерабатывающих предприятиях, руб.; $Е_н$ – нормированный коэффициент эффективности капитальных вложений; V – объем производимой продукции за сезон, тыс. т.

В качестве оценки (c_{ij}) рассматриваемого плана используется сумма приведенных удельных затрат:

$$c_{ij} = \left[\frac{С_{сх} + Е_н К_{сх}}{V}\right]_{ij} + \left[\frac{С_и + Е_н К_{си}}{V}\right]_{ij} + \left[\frac{С_{пп} + Е_н К_{пп}}{V}\right]_{ij},$$

где ij – индекс маршрута между предприятиями.

Среди возможных вариантов находят такой, для которого сумма удельных приведенных затрат по всем маршрутам минимальная.

Для оптимизации схем связей перерабатывающих и сельскохозяйственных предприятий при изменяющихся капитальных вложениях во всех звеньях АПК в качестве оценок плана может быть использована сумма удельных ежегодных издержек на производство конечной продукции:

$$c_{ij} = \left(\frac{С_{сх}}{V}\right)_{ij} + \left(\frac{С_и}{V}\right)_{ij} + \left(\frac{С_{пп}}{V}\right)_{ij}.$$

Пример. В свеклосахарный комплекс республики входят четыре сахарных завода и 10 хозяйств, которые могут быть отнесены к сырьевым зонам этих заводов (табл. 7.20).

Следует определить оптимальный вариант закрепления хозяйств за сахарными заводами, при котором полная себестоимость производства продукта будет минимальной. Для оценки плана с использованием общих затрат на 1 т сахарной свеклы (c_{ij}) производят их расчет по себестоимости, доставке, затратам. Например, себестоимость производства 1 т сахарной свеклы в хозяйстве первом на планируемый период составляет 38 тыс. руб, доставка ее на Скидельский завод 93 тыс. руб, затраты на заводе – 320 тыс. руб. Сумма общих затрат составит 451 тыс. руб. (c_{ij} округляем и уменьшаем до 0,45, см. табл. 7.20). Аналогичным путем рассчитаны и уменьшены на три порядка все c_{ij} в данной таблице для простоты расчета. Сумма общих затрат должна стремиться к минимуму $F = \sum (c_{ij}\, x_{ij}) \rightarrow \min$.

Полученные данные в таблице перераспределяем по методу функционала или потенциала до получения оптимального плана, который представлен в табл. 7.21.

Таблица 7.20

Исходные данные

Хозяйства		Сахарные заводы, затраты на 1 т, тыс. руб. (c_{ij})			
Потребность на заводах, т / Произведено в хозяйстве, т		Скидель	Жабинка	Городея	Слуцк
		3500	1900	1770	1490
1	2300	0,45	0,68	0,40	0,43
2	750	0,60	0,63	0,55	0,60
3	820	0,70	0,52	0,70	0,58
4	500	0,54	0,75	0,65	0,70
5	100	0,55	0,59	0,75	0,52
6	420	0,73	0,60	0,66	0,77
7	540	0,68	0,63	0,45	0,44
8	910	0,56	0,55	0,43	0,62
9	720	0,68	0,67	0,57	0,53
10	700	0,70	0,48	0,50	0,65

Выводы. Значение целевой функции (F) в оптимальном плане 4241,1 тыс. руб. Минимальная себестоимость производства сахара для 10 хозяйств, равная 4241,1 тыс. руб., будет достигнута, если корнеплоды будут поставлять:

на Скидельский завод хозяйства 1 – 2250 т, 2 – 750, 4 – 500 т;

на Жабинковский завод хозяйства 3 – 780 т, 6 – 420, 10 – 700 т;

на Городейский завод хозяйства 5 – 860 т, 8 – 910 т;

на Слуцкий завод хозяйства 1 – 50, 3 – 40, 5 – 140, 7 – 540, 9 – 720 т.

Таблица 7.21

Оптимальный план формирования сырьевых зон перерабатывающих предприятий

Хозяйства и запасы свеклы, т		Сахарные заводы и их потребность, т			
		Скидель **3500**	Жабинка **1900**	Гордея **1770**	Слуцк **1490**
1	**2300**	0,45 **2250**	0,58	0,40	0,43 **50**
2	**750**	0,60 **750**	0,63	0,55	0,60
3	**820**	0,70	0,52 **780**	0,70	0,58 **40**
4	**500**	0,54 **500**	0,75	0,65	0,70
5	**1000**	0,55	0,59	0,45 **860**	0,52 **140**
6	**420**	0,73	0,60 **420**	0,66	0,70
7	**540**	0,65	0,63	0,44	0,44 **540**
8	**910**	0,56	0,55	0,43 **910**	0,62
9	**720**	0,69	0,67	0,57	0,53 **720**
10	**700**	0,70	0,48 **700**	0,50	0,65

Глава 8. ДИНАМИЧЕСКИЕ РЯДЫ

Ряд расположенных в хронологической последовательности значений статистических показателей представляет собой *временной* (*динамический*) *ряд*, анализ которого производится с использованием тренд-анализа.

Статистические показатели, характеризующие изучаемый объект, называют *уровнями ряда*. В динамическом ряду они могут быть *абсолютными, относительными или средними величинами.* Ряды динамики, представленные за определенный промежуток времени, называются *интервальными*. В результате суммирования уровней интервального динамического ряда получаем *накопленные итоги.* Вследствие многих обстоятельств однородность величин, составляющих динамический ряд, может нарушаться, и таким образом изменяется сопоставимость уровней динамического ряда. Если каждый уровень динамического ряда сравнивается с одним и тем же предшествующим первоначальным уровнем, то это сравнение считается с *первоначальной базой*, при сравнении каждого уровня с предшествующим уровнем – с *переменной базой*.

Для представления модели динамического ряда используется *аналитическое выравнивание ряда динамики*. Закономерно изменяющийся уровень изучаемого показателя оценивается как функция времени. В табл. 8.1 приводятся различные виды трендовых моделей, наиболее часто используемые для аналитического выравнивания.

Таблица 8.1

Виды трендовых моделей

Название функции	Описание функции
Линейная	$\hat{Y}_t = b_0 + b_1 t$
Парабола второго порядка	$\hat{Y}_t = b_0 + b_1 t + b_2 t^2$
Кубическая парабола	$\hat{Y}_t = b_0 + b_1 t + b_2 t^2 + b_3 t^3$
Показательная	$\hat{Y}_t = b_0 \cdot b_1 t$
Экспоненциальная	$\hat{Y}_t = b_0 \cdot e b_1^t$
Модифицированная экспонента	$\hat{Y}_t = b_0 + b_1 \cdot b_2^t$
Кривая Гомперца	$\hat{Y}_t = b_0 \cdot b b_1^t$
Логистическая кривая	$\hat{Y}_t = \dfrac{b_0}{1 + b_1 e^{-b_2 t}}$
Логарифмическая парабола	$\hat{Y}_t = b_0 b_1^t b_2^{t^2}$
Гиперболическая	$\hat{Y}_t = b_0 + b_1 \cdot (1 / t)$

Выбор формы кривой определяет результаты экстраполяции тренда. Одним из наиболее распространенных приемов сглаживания уровней с первоначального ряда динамики – это *метод скользящей средней*.

Выполнить прогноз по уравнению тренда можно путем экстраполяции тенденции, наблюдавшейся в прошлом. Уровень динамического ряда ($\hat{y}$), полученный в результате экстраполяции, используется для установления прогноза.

Наличие зависимости между последующими и предшествующими уровнями динамического ряда называют *автокорреляцией*, а построение модели зависимости будущих значений рассматриваемого показателя от прошлых его значений называется *авторегрессией.*

Ряд исследований проводятся длительное время (мониторинг), чтобы выявить тенденцию или закономерность развития и прогнозирования какого-либо процесса или явления. С помощью тренд-анализа описываются характерные тенденции изменения явления во времени, подбираются статистические модели, описывающие эти изменения, производится поиск промежуточных значений путем интерполяции, предсказание результатов значений в перспективе (экстраполяция).

Динамические ряды бывают *простые* (описание одного явления), *сложные* (несколько явлений), *производные* (составленные из средних или относительных величин), *моментные* (оценка события за определенный момент времени), *интервальные* (анализ явления за год, полгода, месяц).

Для создания линии тренда по данным диаграммы используется регрессионный анализ, описывающий взаимодействие между переменными. Следует лишь выбрать один из шести способов аппроксимации данных: линейная, логарифмическая, полиномиальная, степенная, экспоненциальная, скользящая средняя.

8.1. ПОКАЗАТЕЛИ ДИНАМИЧЕСКОГО РЯДА

На первом этапе статистической обработки динамических рядов анализируются основные тенденции (*тренд*) изменения явления во времени. Используется графическое изображение, которое дает исчерпывающую информацию. Вычисляется комплекс специальных показателей, позволяющих дать количественную оценку динамики анализируемого явления.

Абсолютный прирост или *убыль* характеризует изменение явления в единицу или интервал времени. Вычитают из данных последующего периода данные предыдущего. Если ряд возрастает, то прирост считается положительным.

Темп роста или снижения – соотношение в процентах последующего уровня к предыдущему и умноженное на 100. Положительный прирост имеет показатель более 100 %, отрицательный – менее 100 %.

Темп прироста показывает, на сколько процентов увеличился или уменьшился уровень явления. Отражает относительную скорость изменения явления от одного отрезка времени к другому. Вычисляется путем деления абсолютного прироста на предыдущий уровень, либо вычитанием из показателя темпа роста 100. При положительном приросте показатель больше нуля, при отрицательном – меньше нуля.

Абсолютное значение 1 % *прироста* характеризует значение или стоимость 1 % прироста изучаемого явления. Может вычисляться делением абсолютного прироста на темп прироста, или делением показателя предыдущего уровня на 100. «Стоимость» 1 % темпа роста и прироста в различных совокупностях разная.

Пример. Число районов г. Минска с высоким уровнем загрязнения атмосферного воздуха в 2004 г. было 4, в 2005 г. стало 8. Темп роста составил 200 %. В г. Новополоцке таких районов в 2004 г. было 10, а в 2005 г. стало 15. Темп роста составил 50 %. Однако в первом случае число неблагополучных районов увеличилось на 4, во втором – на 5. Это говорит о том, что даже в одном динамическом ряду значение 1 % роста и темпа прироста может существенно отличаться на разных отрезках времени.

Показатель наглядности характеризует динамику явления в процентах относительно исходного уровня, который принимается за 100. В отличие от других показателей стоимость одного процента здесь остается неизменной. Однако динамика изменения исходных данных от одного промежутка времени к другому становится менее выразительной.

Существуют различные варианты вычисления показателей динамики. Они отличаются набором исходных данных и трудоемкостью вычислений (табл. 8.2).

Таблица 8.2

Уровень производства промышленной продукции (ПП) предприятия

Год	Уровень ПП	Абсолютный прирост	Темп роста	Темп прироста	1 % прироста	Показатель наглядности
	У	А	Т	Р	П	Н
1985	65,8					100,0
1986	90,2	24,4	137,1	37,1	0,7	137,1
1987	67,4	–22,8	74,7	–25,3	0,9	102,1
1988	94,3	26,9	139,9	39,9	9,7	143,3
1989	55,4	–38,9	58,7	–41,3	0,9	84,2
1990	45,1	–10,3	81,4	–18,6	0,6	68,5
1991	48,2	3,1	106,9	6,9	0,5	73,3

Приведем примеры расчета показателей, представленных в табл. 8.2. Абсолютный прирост в 1986 и 1987 годах:

$$А_{86} = У_{86} – У_{85} = 90,2 – 658 = 24,4;\ А_{87} = У_{87} – У_{86} = 67,4 – 90,2 = –22,8.$$

Темп роста в 1986 и 1987 годах:

$$Т_{86} = (У_{86} / У_{85}) \cdot 100 = (90,2 / 65,8) \cdot 100 = 137,1;$$

$$Т_{87} = (У_{87} / У_{86}) \cdot 100 = (67,4 / 90,2) \cdot 100 = 74,7.$$

Темп прироста в 1986 и 1987 годах:
первый способ расчета –

$$Р_{86} = (А_{86} / У_{85}) \cdot 100 = (24,4 / 65,8) \cdot 100 = 37,1;$$

$$Р_{87} = (А_{87} / У_{86}) \cdot 100 = (–22,8 / 90,2) \cdot 100 = –5,3;$$

второй способ расчета –

$$Р_{86} = Т_{86} – 100 = 137,1 – 100 = 37,1;$$

$$Р_{87} = Т_{87} – 100 = 74,4 – 100 = –25,3.$$

Абсолютное значение 1 % прироста в 1986 и 1987 годах:
первый способ расчета –

$$П_{86} = У_{85} / 100 = 65,8 / 100 = 0,66;$$

$$П_{87} = У_{86} / 100 = 90,2 / 100 = 0,9;$$

второй способ расчета – $П_{86} = А_{86} / Р_{86} = 24,4 / 37,1 = 0,7;$

$$П_{87} = А_{87} / Р_{87} = –22,8 / –25,3 = 0,9.$$

Показатель наглядности прироста в1986, 1987 г. по сравнению с 1985 г.:

$$Н_{86} = (У_{86} / У_{85}) \cdot 100 = (90,2 / 65,8) \cdot 100 = 137,1;$$

$$Н_{87} = (У_{87} / У_{85}) \cdot 100 = (67,4 / 65,8) \cdot 100 = 102,4.$$

Вычисление средних. Расчет средней в моментном ряду с равными промежутками между датами:

$$М = (1/2\ У_{85} + У_{86} + У_{87} + \ldots + 1/2\ У_{91}) / n,$$

где n – число анализируемых наблюдений.

Средний уровень в моментном ряду с неравными промежутками между датами:

$$М = (1/2\ У_{85} \cdot t_{85} + У_{86} \cdot t_{86} + \ldots + 1/2\ У_{91} \cdot t_{91}) / (t_{85} + t_{86} + \ldots + t_{91}),$$

где t – число дней в году.

Средний уровень в интервальном ряду: $М = (У_{85} + У_{86} + \ldots + У_{91}) / n$.

Средний абсолютный прирост: $М = (А_{85} + А_{86} + \ldots + А_{91}) / n$.

Средний темп прироста (среднее хронологическое) вычисляется в виде среднего геометрического: $М_г = \sqrt[n]{P_{85} \cdot P_{86} \cdot \ldots P_{91}}$.

Динамический характер всех используемых показателей может принимать самые разнообразные формы. Например, абсолютные приросты могут быть стабильными, а темпы роста (прироста) при этом увеличиваться или уменьшаться.

8.2. СГЛАЖИВАНИЕ ДИНАМИЧЕСКИХ РЯДОВ

Углубленный анализ временных рядов требует использования более сложных методик математической статистики. При наличии в динамических рядах значительной случайной ошибки (шума) применяют один из двух простых приемов – *сглаживание* или *выравнивание* путем укрупнения интервалов и вычисления групповых средних. Этот метод позволяет повысить наглядность ряда, если большинство «шумовых» составляющих находятся внутри интервалов. Однако, если «шум» не согласуется с периодичностью, распределение уровней показателей становится грубым, что ограничивает возможности детального анализа изменения явления во времени.

Более точные характеристики получаются, если используют *скользящие средние* – широко применяемый способ для сглаживания показателей среднего ряда. Он основан на переходе от начальных значений ряда к средним в определенном интервале времени. В этом случае интервал времени при вычислении каждого последующего показателя как бы скользит по временному ряду.

Применение скользящего среднего полезно при неопределенных тенденциях динамического ряда или при сильном воздействии на показатели циклически повторяющихся выбросов (резко выделяющиеся варианты или интервенция).

Чем больше интервал сглаживания, тем более плавный вид имеет диаграмма скользящих средних. При выборе величины интервала сглаживания необходимо исходить из величины динамического ряда и содержательного смысла отражаемой динамики. Большая величина динамического ряда с большим числом исходных точек позволяет использовать более крупные временные интервалы сглаживания (5, 7, 10 и т. д.). Если процедура скользящего среднего используется для сглаживания не сезонного ряда, то чаще всего величину интервала сглаживания принимают равной 3 или 5.

Приведем пример вычисления скользящего среднего числа хозяйств с высокой урожайностью (более 30 ц/га) (табл. 8.3).

Таблица 8.3

Сглаживание динамического ряда укрупнением интервалов и скользящим средним

Учетный год	Число хозяйств с высокой урожайностью	Суммы за три года	Скользящие за три года	Скользящие средние
1982	84			90,0
1983	94	270	90,0	89,7
1984	92			88,7
1985	83			87,3
1986	91	262	87,3	87,0
1987	88			86,7
1988	82			83,0
1989	90	249	83,0	82,3
1990	77			82,3
1991	80			82,6
1992	90	248	82,7	82,7
1993	78			

Примеры вычисления скользящего среднего:

1982 г. (84 + 94 + 92) / 3 = 90,0;
1983 г. (94 + 92 + 83) / 3 = 89,7;
1984 г. (92 + 83 + 91) / 3 = 88,7;
1985 г. (83 + 91 + 88) / 3 = 87,3.

Составляется график. На оси абсцисс указываются годы, на оси ординат – число хозяйств с высокой урожайностью. Указываются координаты числа хозяйств на графике и соединяются полученные точки ломаной линией. Затем указываются координаты скользящей средней по годам на графике и соединяются точки плавной полужирной линией.

Более сложным и результативным методом является сглаживание (выравнивание) рядов динамики с помощью различных *функций аппроксимации.* Они позволяют формировать плавный уровень общей тенденции и основную ось динамики.

Наиболее эффективным методом сглаживания с помощью математических функций является *простое экспоненциальное сглаживание.* Этим методом учитываются все предшествующие наблюдения ряда по формуле:

$$S_t = \alpha \cdot X_t + (1 - \alpha) \cdot S_{t-1},$$

где S_t – каждое новое сглаживание в момент времени t; S_{t-1} – сглаженное значение в предыдущий момент времени $t - 1$; X_t – фактическое значение ряда в момент времени t; α – параметр сглаживания.

Если α = 1, то предыдущие наблюдения полностью игнорируются; при величине α = 0 игнорируются текущие наблюдения; значения α между 0 и 1 дают промежуточные результаты. Изменяя значения этого параметра, можно подобрать наиболее приемлемый вариант выравнивания. Выбор оптимального значения α осуществляется путем анализа полученных графических изображений исходной и выравненной кривых, либо на основе учета суммы квадратов ошибок (погрешностей) вычисленных точек. Практическое использование этого метода следует проводить с использованием ЭВМ в программе MS Excel. Математическое выражение закономерности динамики данных можно получить с помощью *функции экспоненциального сглаживания*.

8.3. ВЫРАВНИВАНИЕ ПО СПОСОБУ НАИМЕНЬШИХ КВАДРАТОВ

Предлагаемый способ один из самых эффективных. Суть его следующая: из бесконечного числа линий, которые могли бы быть теоретически проведены между точками, изображающими исходный ряд, выбирается только одна прямая, которая имела бы наименьшую сумму квадратов отклонений исходных (эмпирических) точек от этой теоретической прямой. Выравнивание проводят по уравнению прямой $y = a + bt$, или по уравнению параболы второго порядка $y = a + bt + ct^2$. В основе выбора параболы для выравнивания лежит предположение о том, что не скорость динамики, а ускорение является постоянной величиной. В качестве постоянных величин выступают a, b, c порядкового номера какого-либо периода – t. После расчета постоянных величин a и b известным способом получаем следующее уравнение прямой, по которому вычисляем ряд выравнивания y^1 (табл. 8.4):

$$y^1 = 18{,}748 + 1{,}8382\ t;\ R^2 = 0{,}4047.$$

Показателем правильности выбора того или иного уравнения служит коэффициент R^2. Чем ближе его значение к единице, тем больше соответствие фактического и выравненного распределений.

Современные программы статистической обработки позволяют получать различные теоретические кривые в автоматическом режиме. По результатам можно проводить экстраполяцию или интерполяцию рядов.

Пример. Дать прогноз на следующий шестнадцатый год (см. табл. 8.4) с использованием уравнения регрессии: $У_{16} = 18{,}768 + 1{,}832 \cdot 16 = 48{,}06$.

Таблица 8.4

Выравнивание динамического ряда по способу наименьших квадратов

Номер года	Фактический уровень	Отклонение от центра	Расчетные параметры уравнений	Произведение *yd*	Ряд выравнивания
t	*y*	*d*	d^2	*yd*	y^1
1	16,5	–7	49	–115,5	20,6
2	14,3	–6	36	–85,8	22,4
3	44,0	–5	25	–220,0	24.3
4	35,6	–4	16	–142,4	26,1
5	30,4	–3	9	–91,2	27,9
6	32,4	–2	4	–64,8	29,8
7	22,5	–1	1	–22,5	31,6
8	28,8	0	0	0	33,5
9	15,2	1	1	15,2	35,3
10	42,0	2	4	84,0	37,1
11	26,6	3	9	79,8	39,0
12	42,6	4	16	170,4	40,8
13	51,3	5	25	256.5	42,6
14	46,2	6	36	277,2	44,5
15	53,4	7	49	373,8	46,3
Итого	501,8		280,0	514,7	

Достоверность статистического прогноза зависит от степени интеграции взаимосвязи явлений, которая обеспечивает сохранение механизма формирования явления и инерционность характера динамики (темп, направление, устойчивость) на протяжении длительного времени. Экстраполяция на очень большой период времени вперед или назад резко снижает точность прогноза при R^2 меньше 0,6.

ЛИТЕРАТУРА

Основная

1. *Калинина, В. Н.* Математическая статистика : учебник. 4-е изд., испр. / В. Н .Калинина, В. Ф. Панкин. – М. : Дрофа, 2002.

2. *Колеснев, В. И*. Экономико-математические методы и моделирование в землеустройстве. Практикум : учеб. пособие / В. И. Колеснев, И. В. Шафранская. – Минск : ИВЦ Минфина, 2007.

3. *Михеева, В. С.* Математические методы в экономической географии : в 2-х ч. Ч. 1 : Применение методов линейного программирования : курс лекций / В. С. Михеева. – М. : изд-во Моск. ун-та, 1981.

4. *Пузаченко, Ю. Г.* Математические методы в экологических и географических исследованиях: учеб. пособие / Ю. Г. Пузаченко. – М. : Академия, 2004.

5. Статистика : учеб. пособие / под ред. М. Р. Ефимовой. – М. : ИНФРА , 2000.

6. *Чертко, Н. К*. Математические методы в физической географии : учеб. пособие для геогр. спец. вузов / Н. К. Чертко. – Минск : изд-во «Университетское», 1987.

7. *Чертко, Н. К*. Математические методы в физической географии : учеб. пособие для геогр. спец. вузов / Н. К. Чертко, А. А. Карпиченко. – Минск : БГУ, 2009.

8. *Шикин, Е. В*. Математические методы и модели в управлении : учеб. пособие / Е. В. Шикин, А. Г. Чхартишвили. – М. : Дело, 2000.

Дополнительная

1. *Боровиков, В. П*. Statistica : Статистический анализ и обработка данных в среде Windows. 2-е изд. / В. П. Боровиков. – М. : информ.-издат. дом «Филинъ», 1998.

2. *Боровиков, В. П.* Программа STATISTICA для студентов и инженеров / В. П. Боровиков. – М. : Компьютер Пресс, 2001.

3. *Сачок, Г. И.* Математико-картографическое моделирование природных условий Белоруссии / Г. И. Сачок, Т. В. Цурканова. – Минск: Наука и техника, 1984.

4. *Тикунов, В. С*. Моделирование в картографии / В. С. Тикунов. – М. : изд-во Моск. ун-та, 1997.

5. *Тюрин, Ю. Н.* Статистический анализ данных на компьютере / Ю. Н. Тюрин, А. А. Макаров. – М. : Инфра М, 1998.

ПРИЛОЖЕНИЯ

Таблица 1

Достаточно большие числа

P	Ошибка опыта *p*, %									
	10	9	8	7	6	5	4	3	2	1
0,75	33	40	51	67	91	132	206	367	827	3308
0,80	41	50	64	83	114	164	256	456	1026	4105
0,85	51	63	80	105	143	207	323	575	1295	5180
0,90	67	83	105	138	187	270	422	751	1690	6763
0,91	71	88	112	146	199	287	449	798	1796	7185
0,92	76	94	119	156	212	306	478	851	1915	7662
0,93	82	101	128	167	227	328	512	911	2051	8207
0,94	88	109	138	180	245	353	552	981	2210	8843
0,95	96	118	150	195	266	384	600	1067	2400	9603
0,96	105	130	164	215	292	421	659	1171	2636	10 544
0,965	111	137	173	226	308	444	694	1234	2778	11 112
0,970	117	145	183	240	327	470	735	1308	2943	11 773
0,975	125	155	196	256	348	502	784	1395	3139	12 559
0,980	135	167	211	276	375	541	845	1503	3382	13 529
0,985	147	182	231	301	410	591	924	1643	3697	14 791
0,990	165	204	259	338	460	663	1036	1843	4146	16 587
0,995	196	243	307	402	547	787	1288	2188	4924	19 698
0,999	270	334	422	552	751	1082	1691	3009	6767	27 069

Таблица 2

Случайные числа

3393	6270	4228	6909	9407	1865	8549	3217	2351	8410
9108	2330	2157	7416	0398	6173	1703	8132	9065	6717
7891	3590	2502	5945	3402	0491	4328	2365	6175	7695
9085	6307	6910	9174	1753	1797	9229	3422	9861	8357
2638	2908	6368	0398	5495	3283	0031	5955	6544	38 383
1313	8338	0623	8600	4950	5414	7131	0134	7241	0651
3897	4202	3814	3505	1599	1649	2784	1994	5775	1406
4380	9543	1646	2815	8415	9120	8062	2421	6161	4634
1618	6309	7909	0874	0401	4301	4517	9197	3350	0434
4858	4676	7363	9141	6133	0549	1972	3461	7116	1496

Окончание табл. 2

5354	9142	0847	5393	5416	6505	7156	5634	9703	6221
0905	6986	9396	3975	9255	0537	2479	4589	0562	5345
1420	0470	8679	2328	3939	1292	0406	5528	3789	2882
3218	9080	6604	1813	8209	7039	2086	3369	4437	3798
9697	8431	4387	0622	6893	8788	2320	9358	5904	9539
0912	4964	0502	9683	4636	2861	2876	1273	7870	2030
4636	7072	4868	0601	3894	7182	8417	2367	7032	1003
2515	4734	9897	6761	5636	2949	3979	8650	3430	0635
5964	0412	5012	2369	6461	0678	3693	2928	3740	8047
7848	1523	7904	1521	1455	7089	8094	9872	0898	7174
5182	2571	3643	0707	3434	6818	5729	8615	4298	4129
8438	8325	9886	1805	0226	2310	3675	5058	2515	2388
8166	6349	0319	5436	6838	2460	6433	0644	7428	8556
9158	8263	6504	2562	1160	1526	1816	9690	1215	9590
6061	3525	4048	0382	4224	7148	8256	6526	5340	4064

Таблица 3

Значение критерия τ в зависимости от объема выборки *N* и уровня значимости α

N	α		*N*	α	
	0,05	0,01		0,05	0,01
4	0,955	0,991	17	0,359	0,460
5	0,807	0,916	18	0,349	0,449
6	0,669	0,805	19	0,341	0,439
7	0,610	0,740	20	0,334	0,430
8	0,544	0,683	21	0,327	0,421
9	0,512	0,635	22	0,320	0,414
10	0,477	0,597	23	0,314	0,407
11	0,450	0,566	24	0,309	0,400
12	0,428	0,541	25	0,304	0,394
13	0,410	0,520	26	0,299	0,389
14	0,395	0,502	27	0,295	0,383
15	0,381	0,486	28	0,291	0,378
16	0,369	0,472	29	0,287	0,374
			30	0,283	0,369

Таблица 4

Значения критерия Стьюдента *t* при различных уровнях значимости

ν	Уровни вероятности			ν	Уровни вероятности		
	0,95	0,99	0,999		0,95	0,99	0,999
2	4,30	9,93	31,60	21	2,08	2,83	3,82
3	3,18	5,84	12,94	22	2,07	2,82	3,79
4	2,78	4,60	8,61	23	2,07	2,81	3,77
5	2,57	4,03	6,86	24	2,06	2,80	3,75
6	2,45	3,71	5,96	25	2,06	2,79	3,73
7	2,37	3,50	5,41	26	2,06	2,78	3,71
8	2,31	3,36	5,04	27	2,05	2,77	3,69
9	2,26	3,25	4,78	28	2,05	2,76	3,67
10	2,23	3,17	4,49	29	2,04	2,76	3,66
11	2,20	3,11	4,44	30	2,04	2,75	3,65
12	2,18	3,06	4,32	40	2,02	2,70	3,55
13	2,16	3,01	4,22	50	2,01	2,68	3,50
14	2,15	2,98	4,14	60	2,00	2,66	3,46
15	2,13	2,95	4,07	80	1,99	2,64	3,42
16	2,12	2,92	4,02	100	1,98	2,63	3,39
17	2,11	2,90	3,97	120	1,98	2,63	3,37
18	2,10	2,88	3,92	200	1,97	2,60	3,34
19	2,09	2,86	3,88	500	1,96	2,59	3,31
20	2,09	2,85	3,85	∞	1,96	2,58	3,29

Таблица 4

Значения критерия хи-квадрат (Пирсона)

Степень свободы,ν	Уровни вероятности *P*		
	0,95	0,99	0,999
1	3,841	6,635	10,827
2	5,991	9,210	13,815
3	7,815	11,345	16,268
4	9,488	13,277	18,465
5	11,070	15,086	20,517
6	12,592	16,812	22,457
7	14,067	18,475	24,322
8	15,507	20,090	26,125
9	16,919	21,666	27,877
10	18,307	23,209	29,588

Окончание табл. 4

11	19,675	24,725	31,264
12	21,026	26,217	32,909
13	22,362	27,688	34,528
14	23,685	29,141	36,123
15	24,996	30,578	37,697
16	26,296	32,000	39,252
17	27,587	33,409	40,790
18	28,869	34,805	42,312
19	30,144	36,191	43,820
20	31,410	37,566	45,315
21	32,671	38,932	46,797
22	33,924	40,289	48,268
23	35,172	41,638	49,728
24	36,415	42,980	51,179
25	37,652	44,314	52,620
26	38,885	45,642	54,052
27	40,113	46,963	55,476
28	41,337	48,278	56,893
29	42,557	49,588	58,302
30	43,773	50,892	59,703

Таблица 5

Критические значения F (критерия Фишера)

ν_2*	ν_1 – степени свободы для большей дисперсии																			
	3	4	5	6	7	8	9	10	12	14	16	20	30	40	50	75	100	200	500	∞
3	9,28 26,46	9,12 28,71	9,01 28,24	8,94 27,91	8,88 27,67	8,84 27,34	8,81 27,34	8,78 27,23	8,74 27,05	8,71 26,92	8,69 26,83	8,66 26,69	8,62 26,50	8,60 26,41	8,58 26,35	8,57 26,27	8,56 26,23	8,54 26,18	8,54 26,14	8,53 26,12
4	6,59 16,69	6,39 15,98	6,26 15,52	6,16 15,21	6,09 14,98	6,00 14,66	5,96 14,54	5,96 14,54	5,91 14,37	5,87 14,24	5,84 14,15	5,80 14,02	5,74 13,83	5,71 13,74	5,70 13,69	5,68 13,61	5,66 13,57	5,65 13,52	5,64 13,48	5,63 13,46
5	5,41 12,06	5,19 11,39	5,05 10,97	4,95 10,67	4,88 10,45	4,78 10,15	4,71 10,05	4,74 10,05	4,68 9,89	4,64 9,77	4,60 9,68	4,56 9,55	4,50 9,38	4,46 9,29	4,44 9,24	4,42 9,17	4,40 9,13	4,38 9,07	4,37 9,04	4,37 9,02
6	4,76 9,78	4,53 9,15	4,39 8,75	4,28 8,47	4,21 8,26	4,10 7,98	4,06 7,87	4,06 7,87	4,00 7,72	3,96 7,60	3,92 7,52	3,87 7,39	3,81 7,23	3,77 7,14	3,75 7,09	3,72 7,02	3,71 6,99	3,69 6,94	3,68 6,90	3,67 6,88
7	4,35 8,45	4,12 7,85	3,97 7,46	3,87 7,19	3,79 7,00	3,68 6,71	3,63 6,62	3,63 6,62	3,57 6,47	3,52 6,35	3,49 6,27	3,44 6,07	3,38 5,90	3,34 5,85	3,32 5,78	3,29 5,75	3,28 5,70	3,25 5,67	3,24 5,66	3,23 5,65
8	4,07 7,59	3,84 7,01	3,69 6,63	3,58 6,37	3,50 6,19	3,39 5,91	3,34 5,82	3,34 5,82	3,28 5,67	3,23 5,56	3,20 5,48	3,15 5,36	3,08 5,20	3,05 5,11	3,03 5,06	3,00 5,00	2,98 4,96	2,96 4,91	2,94 4,88	2,93 4,86
9	3,86 6,99	3,63 6,42	3,48 6,06	3,37 5,80	3,29 5,62	3,18 5,35	3,13 5,26	3,13 5,26	3,07 5,11	3,02 5,00	2,98 4,92	2,93 4,80	2,86 4,64	2,82 4,56	2,80 4,51	2,77 4,45	2,76 4,41	2,73 4,36	2,72 4,33	2,71 4,31
10	3,71 6,55	3,48 5,99	3,33 5,64	3,22 5,39	3,14 5,21	3,02 4,95	2,97 4,85	2,97 4,85	2,91 4,71	2,86 4,60	2,82 4,52	2,77 4,41	2,70 4,25	2,67 4,17	2,64 4,12	2,62 4,05	2,59 4,01	2,56 3,96	2,55 3,93	2,54 3,91
11	3,59 6,22	3,36 5,67	3,20 5,32	3,09 5,07	3,01 4,88	2,90 4,63	2,86 4,54	2,86 4,54	2,78 4,40	2,74 4,29	2,70 4,21	2,65 4,10	2,57 3,94	2,53 3,86	2,50 3,80	2,47 3,74	2,45 3,70	2,42 3,66	2,41 3,62	2,40 3,60
12	3,49 5,95	3,26 5,41	3,11 5,06	3,00 4,82	2,92 4,65	2,80 4,39	2,76 4,30	2,76 4,30	2,69 4,16	2,64 4,05	2,60 3,98	2,54 3,86	2,46 3,70	2,42 3,61	2,40 3,56	2,36 3,49	2,35 3,46	2,32 3,41	2,31 3,38	2,30 3,36

Окончание табл. 5

ν_2*	ν_1 – степени свободы для большей дисперсии																			
	3	4	5	6	7	8	9	10	12	14	16	20	30	40	50	75	100	200	500	∞
13	3,41 5,74	3,18 5,20	3,02 4,86	2,92 4,62	2,84 4,44	2,72 4,19	2,67 4,10	2,67 4,10	2,60 3,96	2,55 3,85	2,51 3,78	2,46 3,67	2,38 3,51	2,34 3,42	2,32 3,37	2,28 3,30	2,26 3,27	2,24 3,21	2,22 3,18	2,21 3,16
14	3,34 5,56	3,11 5,03	2,96 4,69	2,85 4,46	2,77 4,28	2,65 4,03	2,60 3,94	2,60 3,94	2,53 3,80	2,48 3,70	2,44 3,62	2,39 3,51	2,31 3,34	2,27 3,26	2,24 3,21	2,21 3,14	2,19 3,11	2,16 3,06	2,14 3,02	2,13 3,00
15	3,29 5,42	3,06 4,89	2,90 4,56	2,79 4,32	2,70 4,14	2,59 3,89	2,55 3,80	2,55 3,80	2,48 3,67	2,43 3,56	2,39 3,48	2,33 3,36	2,25 3,20	2,21 3,12	2,18 3,07	2,15 3,00	2,12 2,97	2,10 2,92	2,08 2,89	2,07 2,87
16	3,24 5,29	3,01 4,77	2,85 4,44	2,74 4,20	2,66 4,03	2,54 3,78	2,49 3,69	2,49 3,69	2,42 3,55	2,37 3,45	2,33 3,37	2,28 3,25	2,20 3,10	2,16 3,01	2,13 2,96	2,09 2,89	2,07 2,86	2,04 2,80	2,02 2,77	2,01 2,75
50	2,79 4,20	2,56 3,72	2,40 3,41	2,29 3,18	2,20 3,02	2,07 2,87	2,02 2,70	2,02 2,70	1,95 2,56	1,90 2,46	1,85 2,39	1,78 2,26	1,69 2,10	1,63 2,00	1,60 1,94	1,55 1,86	1,52 1,82	1,48 1,76	1,46 1,71	1,44 1,68
200	2,65 3,88	2,41 3,41	2,26 3,11	2,14 2,90	2,05 2,73	1,92 2,50	1,87 2,41	1,87 2,41	1,80 2,28	1,74 2,17	1,69 2,09	1,62 1,97	1,52 1,79	1,45 1,69	1,42 1,62	1,35 1,53	1,32 1,48	1,26 1,39	1,22 1,33	1,19 1,28
∞	2,60 3,78	2,37 3,32	2,21 3,02	2,09 2,80	2,01 2,64	1,88 2,41	1,83 2,32	1,83 2,32	1,75 2,18	1,69 2,07	1,64 1,99	1,57 1,87	1,46 1,69	1,40 1,59	1,35 1,52	1,28 1,36	1,24 1,36	1,17 1,25	1,11 1,15	1,00 1,09

Примечание. В числителе – для $F_{0,95}$, в знаменателе – для $F_{0,95}$. * Степени свободы для меньшей дисперсии.

Таблица 6

Минимальные существенные значения коэффициентов корреляции

ν	Уровень вероятности (*P*)		ν	Уровень вероятности (*P*)	
	0,95	0,99		0,95	0,99
3	0,94	0,99	26	0,37	0,48
4	0,84	0,93	27	0,37	0,48
5	0,75	0,87	28	0,36	0,46
6	0,71	0,83	29	0,36	0,46
7	0,67	0,80	30	0,35	0,45
8	0,63	0,77	35	0,33	0,42
9	0,60	0,74	40	0,30	0,39
10	0,58	0,71	45	0,29	0,37
11	0,55	0,68	50	0,27	0,36
12	0,53	0,66	60	0,25	0,33
13	0,51	0,64	70	0,23	0,30
14	0,50	0,62	80	0,22	0,28
15	0,48	0,61	90	0,21	0,27
16	0,47	0,59	100	0,20	0,25
17	0,46	0,58	125	0,17	0,23
18	0,44	0,56	150	0,16	0,21
19	0,43	0,56	200	0,14	0,18
20	0,42	0,54	300	0,11	0,15
21	0,41	0,53	400	0,10	0,13
22	0,40	0,52	500	0,09	0,12
23	0,40	0,51	700	0,07	0,10
24	0,39	0,50	900	0,06	0,09
25	0,38	0,49	1000	0,06	0,09

Таблица 7

Соотношение между *r* и *z′* для *z′* значений от 0 до 5*

z′	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0500	0,0599	0,0690	0,0798	0,0898
0,1	0,0997	0,1096	0,1194	0,1293	0,1391	0,1489	0,1587	0,1684	0,1781	0,1878
0,2	0,1974	0,2070	0,2165	0,2260	0,2355	0,2449	0,2543	0,2636	0,2729	0,2821
0,3	0,2913	0,3004	0,3095	0,3185	0,3275	0,3364	0,3452	0,3540	0,3627	0,3714
0,4	0,3800	0,3885	0,3969	0,4053	0,4136	0,4219	0,4301	0,4382	0,4462	0,4542
0,5	0,4621	0,4700	0,4777	0,4854	0,4930	0,5005	0,5080	0,5154	0,5227	0,5299
0,6	0,5370	0,5441	0,5511	0,5581	0,5649	0,5717	0,5784	0,5850	0,5915	0,5980
0,7	0,6044	0,6107	0,6169	0,6231	0,6291	0,6352	0,6411	0,6469	0,6527	0,6584
0,8	0,6640	0,6696	0,6751	0,6805	0,6858	0,6911	0,6963	0,7014	0,7064	07114
0,9	0,7163	0,7211	0,7259	0,7306	0,7352	0,7398	0,7443	0,7487	0,7531	0,7574
1,0	0,7616	0,7658	0,7699	0,7739	0,7779	0,7818	0,7857	0,7895	0,7932	0,7969

Окончание табл. 7

z'	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,1	0,8005	0,8041	0,8076	08110	08144	0,8178	0,8210	0,8243	0,8275	0,8306
1,2	0,8337	0,8367	0,8397	0,8426	0,8455	0,8483	0,8511	0,8538	0,8565	0,8591
1,3	0,8617	0,8643	0,8668	0,8693	0,8717	0,8741	0,8764	0,8787	0,8810	0,8832
1,4	0,8854	0,8875	0,8896	0,8917	0,8937	0,8957	0,8977	0,8996	0,9015	0,9033
1,5	0,9052	0,9069	0,9087	09104	0,9121	0,9138	0,9154	0,9170	0,9186	0,9202
1,6	0,9217	0,9232	0,9246	0,9261	0,9275	0,9289	0,9302	0,9316	0,9329	0,9342
1,7	0,9354	0,9367	0,9379	0,9391	0,9402	0,9414	0,9425	0,9436	0,9447	0,9458
1,8	0,9468	0,9478	0,9498	0,9488	0,9508	0,9518	0,9527	0,9536	0,9545	0,9554
1,9	0,9562	0,9571	0,9579	0,9587	0,9595	0,9603	0,9611	0,9619	0,9626	0,9633
2,0	0,9640	0,9647	0,9654	0,9661	0,9668	0,9674	0,9680	0,9687	0,9693	0,9699
2,1	0,9705	0,9710	0,9716	0,9722	0,9727	0,9732	0,9738	0,9743	0,9748	0,9753
2,2	0,9757	0,9762	0,9767	0,9771	0,9776	0,9780	0,9785	0,9789	0,9793	0,9797
2,3	0,9801	0,9805	0,9809	0,9812	0,9816	0,9820	0,9823	0,9827	0,9830	0,9834
2,4	0,9837	0,9840	0,9843	0,9846	0,9849	0,9852	0,9855	0,9858	0,9861	0,9863
2,5	0,9866	0,9869	0,9871	0,9874	0,9876	0,9879	0,9881	0,9884	0,9886	0,9888
2,6	0,9890	0,9892	0,9895	0,9897	0,9899	0,9901	0,9903	0,9905	0,9906	0,9908
2,7	0,9910	0,9912	0,9914	0,9915	0,9917	0,9919	0,9920	0,9922	0,9923	0,9925
2,8	0,9926	0,9928	0,9929	0,9931	0,9932	0,9933	0,9935	0,9936	0,9937	0,9938
2,9	0,9940	0,9941	0,9942	0,9943	0,9944	0,9945	0,9946	0,9947	0,9949	0,9950
3,0	0,9951									
4,0	0,9993									
5,0	0,9999									

Примечание. Цифры таблицы являются значениями коэффициента корреляции r, соответствующими значениям z', указанным слева и сверху таблицы.

Таблица 8

Значения коэффициента корреляции рангов Спирмена для двусторонних пределов уровня значимости α

n \ α	0,20	0,10	0,05	0,02	0,01	0,002
4	0,8000	0,8000				
5	0,7000	0,8000	0,9000	0,9000		
6	0,6000	0,7714	0,8286	0,8857	0,9429	
7	0,5357	0,6786	0,7450	0,8571	0,8929	0,9643
8	0,5000	0,6190	0,7143	0,8095	0,8571	0,9286
9	0,4667	0,5833	0,6833	0,7667	0,8167	0,9000
10	0,4424	0,5515	0,6364	0,7333	0,7818	0,8667
11	0,4182	0,5273	0,6091	0,7000	0,7455	0,8364

Окончание табл. 8

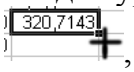
n \ α	0,20	0,10	0,05	0,02	0,01	0,002
12	0,3986	0,4965	0,5804	0,6713	0,7273	0,8182
13	0,3791	0,4780	0,5549	0,6429	0,6978	0,7912
14	0,3626	0,4593	0,5341	0,6220	0,6747	0,7670
15	0,3500	0,4429	0,5179	0,6000	0,6536	0,7464
16	0,3382	0,4265	0,5000	0,5824	0,6324	0,7265
17	0,3260	0,4118	0,4853	0,5637	0,6152	0,7083
18	0,3148	0,3994	0,4716	0,5480	0,5975	0,6904
19	0,3070	0,3895	0,4579	0,5333	0,5825	0,6737
20	0,2977	0,3789	0,4451	0,5203	0,5684	0,6586
21	0,2909	0,3688	0,4351	0,5078	0,5545	0,6455
22	0,2829	0,3597	0,4241	0,4963	0,5426	0,6318
23	0,2767	0,3518	0,4150	0,4852	0,5306	0,6186
24	0,2704	0,3435	0,4061	0,4748	0,5200	0,6070
25	0,2646	0,3362	0,3977	0,4654	0,5100	0,5962
26	0,2588	0,3299	0,3894	0,4564	0,5002	0,5856
27	0,2540	0,3236	0,3822	0,4481	0,4915	0,5757
28	0,2490	0,3175	0,3749	0,4401	0,4828	0,5660
29	0,2443	0,3113	0,3685	0,4320	0,4744	0,5567
30	0,2400	0,3059	0,3620	0,4251	0,4665	0,5479

10. Алгоритм вычисления основных показателей описательной статистики и критерия Стьюдента в Microsoft Office Excel 2003

Решение рассмотрим на примере двух выборок, в которых приведены площади фермерских хозяйств в Брестской и Гомельской областях. Первоначально набираем в ячейках A2–A3 названия областей, в B2–H2 и B3–H3 цифры площадей для каждой области (рис. 1).

	A	B	C	D	E	F	G
1	Площадь фермерских хозяйств						
2	Брестская обл.	300	305	315	320	330	335
3	Гомельская обл.	180	175	190	185	187	197

Рис. 1. Исходные данные для расчетов

Основными статистическими показателями, характеризующими данные выборки, являются: *среднее* (*M*)*, медиану* (Me)*, наименьшее, наибольшее, коэффициент вариации* (*V*)*, среднеквадратическое отклонение* (σ)*, дисперсия* (σ^2). Среднее (*M*) находится следующим образом: выполняем команду *Функция* из меню *Вставка* (или нажимаем на иконку $\mathbf{f_x}$ на панели инструментов), далее в категориях *Статистические* выбираем функцию СРЗНАЧ (рис. 2), сворачиваем появившееся окно нажатием на кнопку  напротив поля **Число 1.** Выделяем ячейки со значениями площадей для первой области (B2 : H2), разворачиваем окно, нажав на эту же кнопку, и жмем [ОК]. Для второй области можно не делать описанную выше процедуру, а воспользоваться функцией автозаполнения: выделяем ячейку с найденным средним значением для первой области (I2), и наведя курсор на правый край клетки I2 до превращения курсора в «крестик»: 

, удерживая левую кнопку мыши, растягиваем выделение на нижележащую клетку (I3), в которой появится значение для второй области.

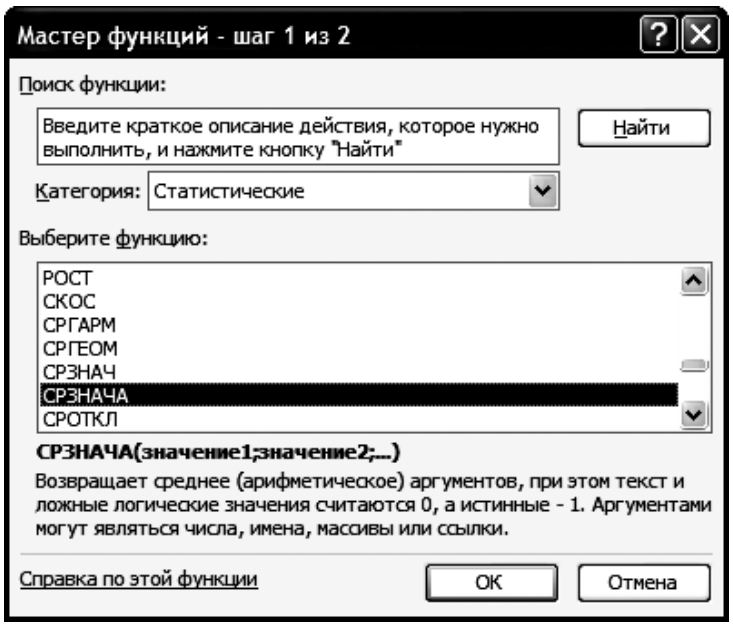


Рис. 2. Окно выбора функции

Аналогичным способом находим медиану (команда МЕДИАНА(B2:H2)), наименьшее = МИН(B2:H2) и наибольшее = МАКС(B2:H2) значения, коэффициент вариации = СТАНДОТКЛОН(B2 : H2)/СРЗНАЧ(B2 : H2)*100, среднеквадратическое отклонение = СТАНДОТКЛОН(B2:H2) и дисперсию = ДИСП(B2:H2). При помощи автозаполнения производим расчет для второй области. MS Excel может производить вычисления при наборе функции вручную, при этом следует помнить, что команды набираются на русском языке, а буквенные обозначения ячеек – латинские.

Расчет базовых статистических показателей может производиться с использованием надстройки (опции) «Пакет анализа», которая позволяет оперативно получить значения показателей описательной статистики. По умолчанию эта опция не установлена, поэтому для ее активации необходимо с помощью команды *Надстройки* из меню *Сервис* открыть окно диалога «*Надстройки*» и в нем установить флажок для компоненты «*Пакет анализа*». После нажатия кнопки [ОК] меню *Сервис* будет дополнено командой *Анализ данных*.

Для расчета показателей выполняем последовательность команду *Анализ данных* из меню *Сервис* в диалоговом окне *Анализ данных* выбираем *Описательная статистика*, в поле «*Входной интервал*» указываем наш (клетки A2 : H3), в поле группирование выбираем «*по строкам*», ставим галочку у «*Метки в первом столбце*» в «*Параметрах вывода*» выбираем «*Выходной интервал*» и указываем там ячейку B5 или другую свободную, отмечаем параметры «*Итоговая статистика*» и «*Уровень надежности*» (значение можно изменять, в нашем случае указываем 95 %), нажимаем [OK].

Нахождение сходства или отличия между двумя выборками с помощью t-теста (критерия Стьюдента). Выбор конкретной команды зависит от типа выборки (зависимая/независимая) и от величин дисперсий. Так, для **независимой** выборки с **различными дисперсиями** выполняются следующие действия: *Сервис – Анализ данных – Двухвыборочный t-тест с различными дисперсиями – ОК.* Для **независимой** выборки с **одинаковыми дисперсиями** алгоритм следующий: *Сервис – Анализ данных – Двухвыборочный t-тест с одинаковыми дисперсиями – ОК*, для **сопряженной** выборки: *Сервис – Анализ данных – Парный двухвыборочный t-тест для средних – ОК*.

В поле «*интервал переменной* 1» указываем интервал значений для первой области (A2 : H2), в поле «*интервал переменной* 2» – интервал значений для второй области (A3 : H3), ставим галочку у окна «*Метки*», далее выбираем «*Выходной интервал*» и указываем там ячейку G5 (или другую свободную), нажимаем [ОК].

В полученных данных *df* – число степеней свободы; *t-статистика* – критерий Стьюдента (фактический); *t критическое двухстороннее* – критерий Стьюдента (табличный). На основании соотношения *t-статистики* (берется по модулю) и *t-критического двухстороннего* делается вывод об достоверности различия выборок.

11. Алгоритм проведения однофакторного дисперсионного анализа в Microsoft Office Excel 2003

Рассмотрим с помощью дисперсионного анализа влияние внесения удобрений на урожайность сельскохозяйственных культур по различным вариантам опыта. В MS Excel набираем исходные данные из индивидуального задания по образцу, показанному на рис. 3:

	A	B	C	D	E
1	Влияние удобрений на урожай с/х культур			Повторности	
2	Варианты	I	II	III	IV
3	фон+100	30,9	28	30,1	28,4
4	фон+200	33,9	36,3	34,3	36
5	фон+300	44,6	44,3	45,6	46,6
6	фон+400	51,1	48,8	50,4	51,4

Рис. 3. Исходные данные

Для анализа используем надстройку «Пакет анализа». Для ее активации необходимо с помощью команды *Надстройки* из меню *Сервис* открыть окно диалога «*Надстройки*» и в нем установить флажок для компоненты «*Пакет анализа*». После нажатия кнопки [ОК] меню *Сервис* будет дополнено командой *Анализ данных* (если надстройка вызывалась ранее и не отключалась, то этот пункт можно пропустить).

Для расчета показателей выполняем последовательность команд *Сервис – Анализ данных – Однофакторный дисперсионный анализ – ОК,* в поле «*Входной интервал*» указываем наш интервал (A3:E6 для приведенного примера), ставим галочки напротив показателей *по строкам* и *метки в первом столбце*; в «*Параметрах вывода*» выбираем «*Выходной интервал*» и указываем там ячейку на этом же листе, значение *Альфа* оставляем прежним, равным 0,05, нажимаем [ОК].

Результаты дисперсионного анализа будут состоять из двух таблиц. В первой таблице для каждого столбца исходной таблицы, в которых располагаются анализируемые группы, приведены числовые параметры: количество чисел (счет), суммы по строкам, средние дисперсии по строкам. Во второй части результатов *MS Excel* использует следующие обозначения: *SS* – сумма квадратов; *df* – степени свободы; *MS* – средний квадрат (дисперсия); *F* – F-статистика Фишера (фактическое значение); *P-значение* – значимость критерия Фишера (критерий является значимым, если величина данного параметра менее 0,05); *F критическое* – критическое (табличное) значение F-статистики при P = 0,05. Путем сравнения *F* и *F критического* делаем вывод. Для данного примера эти значения будут соответственно 252,646 и 3,490295, поэтому влияние удобрений на урожайность доказано.

Если сделать дисперсионный анализ для повторностей опыта (действия аналогичны первоначальным, только вместо показателя *по строкам* выставляется значение *по столбцам* и интервал меняется на B2:E6). В данном случае будет *F* < *F критического*, что и ожидалось, поскольку изменения фактора внутри повторности не происходило.

12. Алгоритм проведения корреляционного и регрессионного анализов в Microsoft Office Excel 2003

Проверим зависимость между баллом пашни (x) и урожайностью многолетних трав (y), для чего набираем в ячейках A2 : K3 следующие данные:

x	43	42	38	36	33	45	40	45	36	32
y	33,2	18,6	28,4	26,5	30,9	31,8	32,4	30,6	26,8	24,4

Строим точечную диаграмму: выделяем набранную таблицу (ячейки A2 : K3) и жмем на пиктограмму  на панели инструментов или *Вставка – Диаграмма,* в закладке *Стандартные* выбираем *Точечная* и первый сверху из имеющихся примеров жмем *Далее,* в закладке *Диапазон данных* отмечаем *Ряды в строках – Далее.* В закладке *Заголовки* в окошке *Ось X* (*категорий*) набираем «Балл пашни» (может отличаться для различных индивидуальных заданий, в этом случае пишется название первого сравниваемого параметра), в окошке *Ось Y* (*значений*) «Урожайность многолетних трав», в закладке *Легенда* снимаем галочку с показателя *«Добавить легенду»* – *Далее – Поместить диаграмму на имеющемся листе – Готово*.

Добавляем линию тренда, для чего кликаем на маркере точки данных правой клавишей и выбираем пункт *Добавить линию тренда* (см. рис. 4).

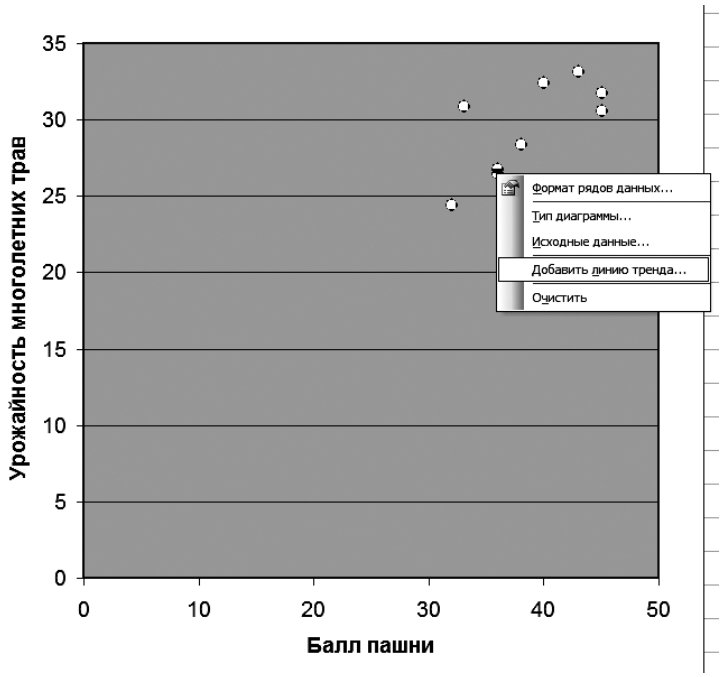


Рис. 4

В закладке *Тип* выбирается *Линейная,* в закладке *Параметры* отмечаются пункты *показывать уравнение на диаграмме* и *поместить на диаграмму величину достоверной аппроксимации – ОК*. В итоге будет построена линия тренда и составлено уравнение линейной регрессии. Находим артефакты – значения, которые сильно отдалены от линии тренда и не вписываются в общую картину (рис. 5). Более правильно выявлять артефакт через п приведенные в п. 1.2. Удаляем эти значения из таблицы данных (в указанном примере очищаются от цифр ячейки С2 : С3), MS Excel автоматически пересчитает уравнение регрессии. В некоторых случаях (при нелинейной корреляции), можно использовать другие варианты линий тренда, например логарифмическую, степенную или экспоненциальную.

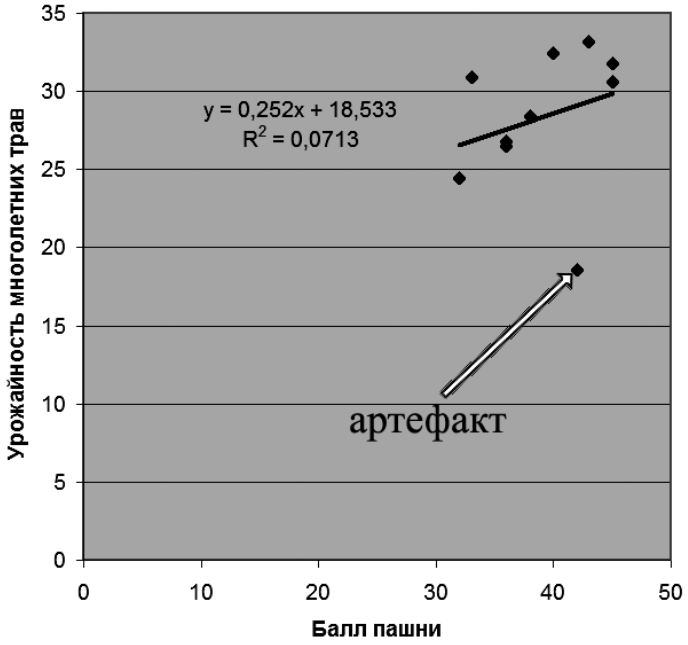


Рис. 5

Рассчитываем коэффициент корреляции, установив курсор в клетку В5, используя команду КОРРЕЛ: *Вставка – Функция* (или иконка $\mathbf{f_x}$ на панели инструментов) – выбираем в категориях *Статистические* – функцию КОРРЕЛ – сворачиваем появившееся окно нажатием на кнопку  напротив поля *Массив 1*. Выделяем ячейки со значениями ***x*** (В2 : К2), далее в поле *Массив 2* выделяем ячейки со значениями ***y*** (В3 : К3), разворачиваем окно, нажав на эту же кнопку, и жмем *ОК*.

Оцениваем значимость коэффициента корреляции (***r***) по критерию Стьюдента по формуле $t_r = \sqrt{N-2}/\sqrt{1-r^2}$ и сравниваем с табличным (критическим) значением, если фактическое значение больше критического, то корреляционная связь существенна, если меньше – недостоверна (вид формул на рис. 6).

	A	B
1	Между балом пашни и урожайностью	многолетних трав
2	x	43
3	y	33,2
4		
5	Коэф. корреляции	=КОРРЕЛ(B2:K2;B3:K3)
6	Критерий Стьюдента	=B5*КОРЕНЬ(СЧЁТ(B2:K2)-2)/КОРЕНЬ(1-B5*B5)
7	Критическое значение критерия Стьюдента	=СТЬЮДРАСПОБР(0,05;СЧЁТ(B2:K2)-2)

Рис. 6

Регрессионный анализ проводится с помощью надстройки «*Пакет анализа*», для последовательности команд *Сервис – Анализ данных – Регрессия,* в поле «*Входной интервал*» указываем значения для Y и X (A3 : K3 и A2 : K2 соответственно), в «*Параметрах вывода*» выбираем «*Выходной интервал*» и указываем там ячейку на этом же листе, отмечаем параметры «*Уровень надежности*» (значение можно изменять, в нашем случае указываем 95 %), нажимаем [OK]. Если удалялся артефакт, то необходимо скопировать первоначальные значения в другие ячейки, поскольку значения во входном интервале должны быть непрерывными.

13. Алгоритм проведения кластерного анализа в Statsoft Statistica 6.0

Проведем кластерный анализ для областей Беларуси по показателям внесения удобрений и урожайности ряда сельскохозяйственных культур.

Допускается выполнение работы по двум вариантам (на выбор пользователя):

а) Набор исходных данных в MS Excel. Открыть MS Excel. Набрать следующие исходные данные в ячейках диапазона A1 : F6 Листа 1. Сохранить введенные данные и закрыть файл.

16,6	212	27,3	175	38,9	12,4
13,9	193	15,7	156	40,1	11,8
16,3	226	25,3	186	28,6	13,9
13,5	240	29,4	178	43,5	15,4
11,6	205	25,9	193	33,6	10,3
15,5	231	27,5	185	32,5	14,4

Запустить программу **Statistica** (через *Пуск – Все программы* или ярлык на рабочем столе), открыть в ней набранный в Excel файл (*File – Open* или через пиктограмму на панели инструментов, в появившемся окне укажите путь к файлу с вышеуказанной таблицей, не забудьте выбрать в окне *«Тип файлов» Excel files (.xls)*). Далее в появившемся диалоговом окне выбираем *Import selected sheet to a Spreadsheet*, затем в следующем окне выбираем *Лист 1 – ОК*, в следующем окне ничего не изменяем и сразу жмем *ОК*.

б) Подобную таблицу можно сразу создать путем набора в программе **Statistica**, пример **а** показывает на возможность импорта данных из MS Excel.

Переименовать в **Statistica** сроки последовательно в *Брестская, Витебская, Гомельская, Гродненская, Могилевская, Минская*, для чего нужно дважды кликнуть на них левой клавишей мышки, а столбцы (Var 1, Var 2 и т. д.) в поле *Name* после двойного щелчка левой клавиши мыши соответственно набираем: *органич. удобр., т/га; минерал. удобр., кг/га; зерновые, ц/га; картофель, ц/га; кормовые травы, ц/га; зернобобовые, ц/га*.

Проводим кластерный анализ, для чего выполняем следующие действия: *Statistics – Multivariate Exploratory Techniques – Cluster analysis – Joining tree clustering* (оно выбрано по умолчанию) – *OK*. В следующем диалоговом окне выбираем закладку *Advanced* – жмем на кнопку *Variables,* там отмечаем все переменные (выделяем левой клавишей мыши при нажатой клавише *Shift* или просто кликаем на кнопке *Select All*) – *OK*. В полях *Input file* ставим *Raw data, Kluster – Cases (rows), Amalgamation (linkage) rule – Single Linkage, Distance Measure – Euclidean distances.* Если ваши параметры соответствуют представленным на рис. 7, то жмем *OK*.

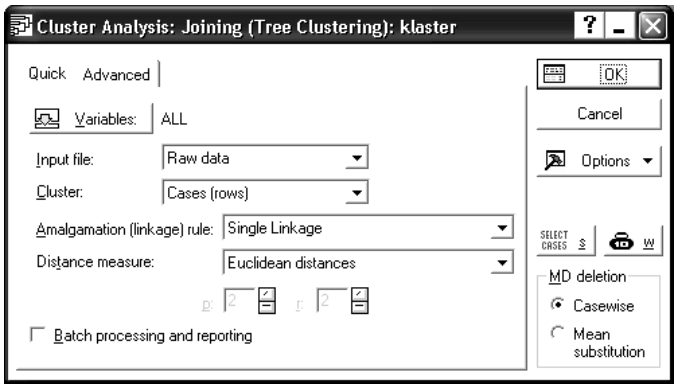


Рис. 7

Далее в появившемся окне нажимаем *Summary.* Появится дендрограмма с разбитием данных на кластеры. После этого нажимаем на кнопку *Joining result:имя файла* (слева в самом низу программного окна). Там, на закладке *Advanced* выбираем по очереди показатели: *Distance matrix, Descriptive statistics* и *Matrix.* Так же можно выбрать вертикальное расположение древа (показатель *Vertical icicle plot*). Полученный график и таблицы используются для интерпретации данных анализа.

14. Алгоритм проведения факторного анализа в Statsoft Statistica 6.0

С помощью факторного анализа оценим плодородие почв в Минском районе под влиянием природных и агротехногенных факторов, для чего набираем в программе **Statistica** следующую таблицу:

8	2,7	3,5	0,3	2,5	4,6	65	21
11	2,9	4,6	0,2	2,4	4,4	63	27
13	3	4,7	0,1	2,3	4,5	64	26
7	1,9	3,1	0,3	2,1	4,5	54	23
8	2,1	3,6	0,4	2,6	4,7	42	24
9	2,3	4,2	0,1	2,7	5,1	43	24
14	2,7	4,1	0,5	2,8	5	60	22
13	2,8	4,5	0,7	2	4,4	52	23
12	2,6	4,7	0,2	2,2	4,5	47	25
14	2,4	4,8	0,4	2,4	4,6	42	26
9	2,2	3,9	0,6	2,6	4,9	51	27

Переименовываем столбцы в **Statistica** (Var 1, Var 2 и т. д.), для чего нужно дважды кликнуть на них левой клавишей мышки и набрать в поле *Name* соответственно: *органические удобрения, т/га; минеральные удобрения, ц/га; дозы извести, т/га; пестициды, кг/га; гумус, т/га; гидролитическая кислотность* (*Н*)*, мэкв/100 г; влажность почвы, %; физическая глина, %.*

Проводим факторный анализ, для чего выполняем следующие действия: *Statistics – Multivariate Exploratory Techniques – Factor analysis – OK*. В следующем диалоговом окне жмем на кнопку *Variables,* там отмечаем все переменные (выделяем левой клавишей мыши при нажатой клавише *Shift* или просто кликаем на кнопке *Select All*) – *OK*. В поле *Input file* ставим *Raw data,* в поле *MD deletion – Casewise* (выставлено по умолчанию) и жмем *OK*.

В следующем окне переходим на закладку *Advanced*, где по умолчанию выбраны *Principal components*, а значение *Max. no. of factors* равно 2. Если выбраны другие значения, то устанавливаем вышеуказанные и жмем *OK*.

В полученном окне, на закладке *Quick* жмем на кнопку *Eigenvalues*. В получившейся таблице *Eigenvalues (Factors)* приведены: 1) собственные значения факторов, которые были выделены; 2) процент объясненной дисперсии; 3) кумулятивные собственные значения и 4) кумулятивный процент объясненной дисперсии. В нашем случае выделилось два фактора.

После этого возвращаемся в диалоговое окно *Factor Analysis Results: factor* (слева в самом низу программного окна), где на закладке *Loadings* выбираем в окне *Factor rotation* показатель *Varimax raw,* после чего нажимаем на кнопку *Summary: Factor loadings* и *Plot of loadings, 2D*. На закладке *Explained Variance* нажимаем по очереди на кнопки: *Scree plot, Communalities.*

Далее переходим на закладку *Descriptives* и нажимаем на кнопку, в новом окне на закладке *Quick* поочередно нажимаем на кнопки *Means & SD* и *Correlations*. Можно вернуться в окно *Factor Analysis Results: factor*, нажав на *Cancel*.

Полученные таблицы и график используются для интерпретации данных анализа.

15. Решение задачи на оптимальность

Требуется обосновать оптимальные размеры отраслей фермерского хозяйства, позволяющие сохранить плодородие пашни и получить максимум прибыли.

Исходная информация.

1. Фермер имеет 40 + К пашни, 1300 + 10 К чел.-дн. годового труда, 600 + 5 К ц единиц кормов с пастбищ и сенокосов.

2. Расход ресурсов и выход продукции на единицу отрасли приведен в таблице.

Показатели	Зерновые	Картофель	Многолетние травы	Коровы	
Площадь пашни, га	1	1	1	–	40 га
Затраты труда, чел.-дн.	10	32	3	24	
Баланс гумуса, т/га	– 0,9	– 1,6	+0,5		
Расход кормов, ц к ед.				50	
Выход кормов, ц к ед.	10	15	25		
Выход навоза от коровы, т				9	
Прибыль, у. д. е.	50 + 0,5 К	60 – 0,5 К		100 – К	

3. Коэффициент перевода органического удобрения в гумус 0,1.

4. Площадь зерновых должна быть не менее 10 га.

Решите задачу симплексным методом и проведите анализ.

Покажем первоначальные условия задачи на листе Excel в виде рабочего листа «Оптимизация».

Решение подобной задачи возможно при помощи надстройки *Поиск решения*, для ее активации необходимо с помощью команды *Надстройки* из меню *Сервис* открыть окно диалога *Надстройки* и в нем установить флажок для компоненты *Поиск решения*. После нажатия кнопки [ОК] меню *Сервис* будет дополнено командой *Поиск решения*.

	A	B	C	D	E	F	G
1	Расчет оптимальной программы развития сельского хозяйства						
2	Показатели	Зерновые	Картофель	Мн. травы	Коровы	Итого	Имеется
3	Площадь, га					=СУММ(B3:D3)	40
4	Поголовье коров, гол						
5	Затраты труда, чел.-дн.	=B3*10	=C3*32	=D3*3	=E4*24	=СУММ(B5:E5)	1300
6	Выход кормов, ц к.ед.	=B3*10	=C3*15	=D3*25		=СУММ(B6:D6)+600	
7	Расход кормов, ц к.ед.				=E4*50	=F4*50	
8	Баланс гумуса, т	=B3*0,9	=C3*1,6	=D3*0,5		=B8+C8-D8	
9	Выход навоза				=E4*9		
10	Прибыль, у.д.е.	=B3*50	=C3*60		=E4*100	=СУММ(B10;C10;E10)	

Рис. 8

В меню *Сервис* Выбираем команду *Поиск решения*. Установить целевую ячейку, которая должна принимать максимальное, минимальное или конкретное значение, в нашем случае это ячейка F10. Ставим отметку тип «максимальное значение». В поле *Изменить ячейки* указываем диапазоны ячеек, оптимальные значения которых требуется найти (B3, C3, D3, E4). Вводим условия ограничения, для чего здесь же вызываем диалоговое окно «ограничение», щелкнув по *Добавить*. В диалоговом окне *Добавление ограничения* в окошке *Ссылка на ячейку* вносим адрес ячейки с функцией *Ограничения*, где указывается число или адрес ячейки, содержащей ограничения (табл. 2). Между ними проставить знаки <= или =>. После ввода всех ограничений выбирают *ОК*.

Появляется диалоговое окно *Поиск решения*, в нем для решения задачи щелкаем по кнопке *Выполнить*. После завершения расчетов появится диалоговое окно *Результаты поиска решений*. В нем помечаем пункт *Сохранить найденное решение* и указываем необходимый тип отчета (*результаты, устойчивости, пределы)*. Далее нажимаем *ОК* для сохранения результата.

Ограничения	Описание
B3 : D3 ≥ 0	Площадь посева сельскохозяйственных культур не может быть отрицательной
E4 ≥ 0	Поголовье коров не может принимать отрицательные значения.
F3 ≤ G3	Общая площадь посева сельхозкультур не должна быть больше площади пашни
F5 ≤ G5	Затраты труда на возделывание сельхозкультур в растениеводстве и животноводстве не могут превышать имеющиеся ресурсы труда
F7 ≤ F6	Расход кормов в животноводстве не должен превышать выхода кормов с отраслей растениеводства с учетом их заготовки на сенокосах и пастбищах
B3 ≥ 10	Площадь зерновых культур должна быть не менее 10 га
F8 ≤ F9	Вынос (минерализация) гумуса с урожаем сельхокультур не должен превышать его поступления с отрасли животноводства

Если решение неверно, то появляется:

- значение целевой ячейки не сходится;
- поиск не может найти подходящее решение;
- условия для линейной модели не удовлетворительны и др.

При положительном решении выбрать *Сохранить сценарий*, при отрицательном – *Восстановить исходные данные*.

Учебное издание

Чертко Николай Константинович

МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ЗЕМЛЕУСТРОЙСТВЕ

Учебно-методическое пособие

Ответственный за выпуск *Е. А. Логвинович*

Дизайн обложки *Т. А. Малько*
Технический редактор *Т. К. Раманович*
Компьютерная верстка *А. А. Микулевича*
Корректор Т. В. Атрошкевич

Электронный ресурс 2 Мб.

Белорусский государственный университет.
Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий № 1/270 от 03.04.2014.
Пр. Независимости, 4, 220030,/Минск.